INERTIA, ANGULARITY, LINEAR TRANSFORMATIONS WITH INVARIANTS AND ITERATIVE SOLUTIONS OF THE LYAPUNOV MATRIX EQUATION

A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

By

C. S. KARUPPAN CHETTY

to the
DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY, KANPUR
AUGUST, 1981

To The Memory of
My kid MEENA

CERTIFICATE

This is to certify that the research work embodied in the present thesis entitled "Inertia, Angularity, Linear Transformations with Invariants and Iterative Solutions of the Lyapunov Matrix Equation" by Mr.C.S.Karuppan Chetty has been carried out under my supervision and that it has not been submitted elsewhere for any degree or diploma.

(R.K.S.RATHORE)

Assistant Professor,
Department of Mathematics,
Indian Institute of Technology,
Kanpur-208016, India.

August, 1981.

ACKNOWLEDGEMENTS

At the very outset, I take this opportunity to express my profound sense of gratitude to my teacher and supervisor, Dr.R.K.S.Rathore for suggesting this topic and for his constant interest, invaluable advice and inspiring guidance throughout the course of this work. His willingness to discuss the material at every stage despite his busy schedule and his ability to provide inspiration and encouragement in moments of doubt, has indeed gone a long way in the successful completion of this thesis.

I wish to express my sincere thanks to the Government of India, for having given me the opportunity of pursuing my further studies at Kanpur under the Quality Improvement Programme. To the Board of Governors, RECT, I am indebted for having sponsored me for this programme. I also acknowledge with thanks the various facilities provided to me by the Indian Institute of Technology, Kanpur during my stay at this Institute.

For moral support and constant encouragement, thanks are due to all my friends, colleagues and relatives. To the Faculty and Staff, Department of Mathematics, IITK I owe my gratitude for their co-operation during the course of my work here.

I think I will fail in my duty if I do not express my word of indebtedness to all those persons who have sent me their work on this subject, through reprints and preprints.

Last, but not the least, my special word of thanks goes to my wife and my daughter Kavitha for their patience and forbearance during this span of three years while I was out at this Institute.

10-8-1981

(C.S.KARUPPAN CHETTY)

CONTENTS

SYNOPSIS

The advent of modern high-speed computers has greatly enhanced the applicability of matrices in all the walks of science and engineering, for instance, in the stability of biological, physical and social systems and control theory. In fact, most of the engineering problems in order to get solved are to be discretized and the corresponding original questions lead to interesting matrix problems. Earlier, the bottleneck was the large size of the corresponding discretized systems and this now gets removed to a large extent as the faster and more versatile computer systems are coming to our aid. In view of this, matrix theory with the influx of varied new problems has now become all the more active and interesting area of research.

The present thesis is concerned with three categories of problems in matrix theory. The first of these concerns the foundations and consists of the development of a notion of angularity which generalizes the well-known concept of inertia due to Ostrowski and Schneider. The second topic arises due to mathematical curiosity as to what are the most general linear transformations on matrices and vectors which possess certain natural invariants. The third and the last topic studied is application-oriented and deals with the iterative numerical solutions of the Lyapunov matrix equation $AX+XA^{*}=C$.

Indeed, the notion of inertia is immediately connected with the stability of a linear system $\dot{x}=Ax$ and a way of determining which is to see whether the corresponding Lyapunov matrix equation mentioned above for a negative definite C leads to a positive definite solution matrix X. Thus the first and the third topics of the present thesis are interconnected. Since in the second topic, we also study certain inertia- and angularity-preserving transformations, the first two topics are as well related.

The thesis consists of four chapters. A chapterwise summary is as follows:

Chapter 1 is introductory in nature and is intended to provide a reasonable up-to-date survey of the three areas closely related with our thesis material. These are inertia theory, linear transformations with invariants and the matrix equation $AX+XB=C$.

In Chapter 2, we introduce and study a notion of angularity of a matrix as a generalization of the well-known concept of inertia. The inertia $In(A)$ of an $n \times n$ complex matrix A is defined as an ordered triple $(\pi(A), \nu(A), \delta(A))$, the entries denoting the total number of eigenvalues of A with positive, negative, and zero real parts respectively. The inertia $In(A)$ thus depends on the distribution of arguments of eigenvalues of A. This dependence becomes complete in the case of angularity $\Theta[A]$ of A. In order to define $\Theta[A]$, we first define the ray space Ω of the complex

plane C following Cain as the set $\{\{0\}\} \cup \{e^{i\theta}\mathbb{R}_+ : 0 \leq \theta < 2\pi\}$, where $\mathbb{R}_+ = (0,\infty)$. A general element of Ω is called a ray and is denoted by ω, $\omega = \{0\}$ being called a null ray and $\omega = e^{i\theta}\mathbb{R}_+$ $(0 \leq \theta < 2\pi)$ being a proper ray. Now the angularity $\theta[A]$ of an $n \times n$ complex matrix A is defined as a mapping from Ω to the set of nonnegative integers for which $\theta[A]_\omega$ is the number of eigenvalues of A (counting multiplicities) lying on the ray ω.

The results of this chapter are mostly related to angularity of normal matrices. Many angularity theorems are proved there and the main result says that if B and C are nonsingular matrices then B^*AB and C^*AC have the same angularity provided they are normal. Some well-known inertia theorems, for example, Sylvester's law of inertia have been deduced as corollaries of this main result. The case when C is permitted to be singular is discussed next. Then the quantitatively sharpened results of Ostrowski and those of Thompson on Sylvester's law have been extended to normal matrices. Further we define totally normal matrices as those having all their principal submatrices as normal and prove that a matrix of order ≥ 3 is totally normal if and only if all its second and third order principal submatrices are normal. Finally, by making use of the main result of this chapter, we prove an angularity result for partitioned normal matrices which states that if A is a normal matrix expressed in the

of the original function such as nonnegativity, monotonicity, number of oscillations(variations), number of zeros etc. are then to be inferred from the corresponding notions about the components of the vectors. From this point of view, it is natural to characterize matrix transformations which preserve such structural characteristics of the vectors. In this connection, we determine matrices which preserve properties such as nonnegativity, variations of various order, number of zeros, nondecreasing trend etc. Finally we determine linear transformations on the set of circulants which preserve inertia and angularity.

The last chapter is motivated by the works of Hoskins, Meek and Walton on the iterative methods for the numerical solution of the Lyapunov matrix equation $AX+XA^*=C$. Even though quite stable and economical direct methods such as the Bartels-Stewart algorithm are available for the problem, due to the surprisingly fast convergence (about 5 iterates) of some of the iterative methods these seem to deserve a more extensive mathematical analysis. In this context, we consider two such iterative procedures and establish their theoretical convergence for certain general classes of matrices.

Also keeping in mind very large, sparse, inconsistent, singular or underdetermined general systems $AX+XB=C$ for which direct methods are often of no avail, we propose compact implementations of projection and residual projection

partitioned form $(A_{ij})_{i,j=1,2}$ with A_{11} and A_{22} being normal, A_{11} nonsingular and $A_{11}^{*}A_{12} = A_{11}A_{21}^{*}$, then

$$\Theta[A]_{\omega} = \Theta[A_{11}]_{\omega} + \Theta[B_{22}]_{\omega}, \text{ for all } \omega \in \Omega$$

where

$$B_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

This in fact generalizes an interesting result due to Haynsworth on the inertia of a partitioned Hermitian matrix.

The next chapter is devoted to the study of linear transformations on matrices and vectors. We first prove that (a) any linear transformation T, on the set of n by n complex matrices, mapping Hermitian matrices into themselves and preserving the inertia of each Hermitian matrix is of the form $T(A) = C^{*}AC$ or $T(A) = C^{*}A'C$ where C is some nonsingular matrix and A' denotes the transpose of A and that (b) any linear transformation T mapping normal matrices into normal matrices and preserving the angularity of each normal matrix is also of one of the above forms, but with $C=kU$ where $k\neq 0$ and U is unitary. Surprisingly, it turns out that a linear transformation T mapping normal matrices into normal matrices preserves inertia of each normal matrix if and only if it preserves the angularity of each normal matrix.

The vectors which arise while discretizing a continuous function at nodal points can be viewed as defining a function on a discrete ordered set. The structural characteristics

methods for such systems, respectively for minimum norm and least squares solutions, which are iterative in nature and do not lead to excessive memory problems. The convergence of these procedures is always guaranteed as it can be easily inferred from the corresponding results for a full system $Ax=b$.

The references in the thesis are separately collected mainly under the three headings corresponding to the three major topics mentioned before. Also, to make the thesis useful for other workers, whenever accessible, along with the most of the references their Mathematical Review and Zbl. abstract numbers have been quoted. Parts of Chapter 2 and Chapter 3 have been accepted for publication in the Journal, Linear Algebra and its Applications.

1. THREE TOPICS IN MATRIX THEORY: A SURVEY

1.1. Introduction

As the title of the thesis indicates, in the chapters to follow we study some problems related to (a) angularity which is a generalization of the notion of inertia, (b) linear transformations having invariant subsets and functionals, and (c) iterative numerical procedures for the solution of the Lyapunov and some more general matrix equations. In the following three sections of this introductory chapter, we have tried to present a reasonably up-to-date survey of the past work having a bearing upon the above three topics. The topics of survey are (i) inertia theory (ii) linear transformations with invariants and (iii) the matrix equation $AX+XB=C$.

The first two topics considered in the thesis derive their importance from their use in the mathematical analysis of various matrix problems while the last topic is important because of its many practical applications which to some extent have been elaborated in the discussion to follow.

Indeed, as is well known, many physical, biological, social and economic systems ultimately have the mathematical description or model defined by the vector-matrix equation

$$\dot{x} = Ax \qquad (1.1.1)$$

where A is a given, constant $n \times n$ matrix and x is an n-dimensional column vector to be determined to satisfy certain initial conditions. In the above equation, $\dot{x}$ denotes the

differentiation of x with respect to time t. The study of stability of such systems has been important in its own right since an unstable system is never acceptable.

It is a well-established fact that the system governed by (1.1.1) is asymptotically stable in the sense that every solution vector $x(t)$ of (1.1.1) approaches zero as $t \to \infty$ if and only if (iff) A is stable, i.e., all the eigenvalues of A have negative real parts (see Bellman [4, p.250]). Conditions for A to be stable can be expressed in terms of the coefficients of the characteristic polynomial of A, the famous one being the Routh-Hurwitz criterion [7, Vol.II, p.194]. For an account of such stability criteria and related problems, one may refer, for instance, Anderson [24], Barnett [1,27,28], Barnett and Šiljak [30], Barnett and Storey [3], Duffin [55], Fuller [56], Gantmacher [7, Vol.II], Householder [62], Lancaster [13], Lehnigk [14], Marden [17], Taussky [84] and Wall [87] and the references given therein.

In practice, these classical approaches of testing the stability suffer from the difficulty of computing accurately the coefficients of the characteristic polynomial. This formidable job of computing these coefficients can very well be avoided if the following well-known result due to Lyapunov is utilized. It says that (see Bellman [4, p.254], Gantmacher [7, Vol.II, p.189]) an $n \times n$ real matrix A is stable iff the matrix equation

$$A^T X + XA = -I, \qquad (1.1.2)$$

A^T denoting the transpose of A and I denoting the $n \times n$

identity matrix, has a symmetric positive definite solution X. The above result can be generalized to a complex matrix A also. Correspondingly, the matrix equation (1.1.2) is replaced by

$$A^*X + XA = - I \tag{1.1.3}$$

where A^* denotes the conjugate transpose of A and in this case the solution matrix X should be Hermitian positive definite. In fact, Lyapunov's theorem holds even if I is replaced by any symmetric (Hermitian) positive definite matrix P in (1.1.2) ((1.1.3)).

Thus the stability problem, namely the problem of knowing whether the matrix A has all its eigenvalues in the open left half plane involves the solution of (1.1.2) or (1.1.3) and once this solution is obtained then its positive definiteness can be tested by the Sylvester determinantal conditions [31], e.g., through Gaussian elimination.

An equation of the form (1.1.2) or (1.1.3) or in general, that of the form

$$A^*X + XA = C \tag{1.1.4}$$

or

$$AX + XA^* = C \tag{1.1.5}$$

where C is a given $n \times n$ complex matrix is called the Lyapunov matrix equation. Apart from determining the stability of A, there are so many applications of solutions of the Lyapunov matrix equation and its more general form

$$AX + XB = C \tag{1.1.6}$$

called the Sylvester equation [232], where A, B, C are given

matrices of order m x m, n x n and m x n respectively and X is an m x n matrix to be solved for. Two such applications are explained below, while several more will be listed in Section 1.4.

In the theory of control processes an important problem [4, 169] is to evaluate the quadratic cost functional of the form

$$J = \int_0^\infty x^T B x \, dt \tag{1.1.7}$$

with x satisfying $\dot{x}=Ax$, $x(0)=c$, A being a real stable matrix of order n. In (1.1.7), x^T denotes the transpose of x and B is a given n x n real matrix. If we suppose

$$x^T B x = - \frac{d}{dt}(x^T Q x) \tag{1.1.8}$$

then we have

$$A^T Q + Q A = - B \tag{1.1.9}$$

which is obtained by expanding the right hand side of (1.1.8) and then substituting $\dot{x} = Ax$. In view of (1.1.8), the integral mentioned in (1.1.7) becomes

$$J = - [x^T Q x]_{t=\infty} + [x^T Q x]_{t=0}. \tag{1.1.10}$$

On the assumption that x tends to zero as $t \to \infty$ and $x(0)=c$, it follows that

$$J = c^T Q c \tag{1.1.11}$$

which can be easily evaluated, once we know Q, the solution of (1.1.9). Thus we see in this process that the solution of the Lyapunov matrix equation enables to avoid the problem of solving the system of differential equations as well as the cumbersome job of evaluating the required improper integral.

The other example illustrates how we can use the solution of the Sylvester equation to solve certain large linear systems arising in boundary value problems. Suppose we wish to solve the two dimensional Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = - f(x,y) \qquad (1.1.12)$$

over a rectangular region covered by a square grid of side h, say with rs interior points arranged in r rows and s columns. Denoting a function value at the grid point (ph,qh) by $u_{p,q}$ we get the system

$$4u_{p,q} - (u_{p-1,q} + u_{p+1,q} + u_{p,q-1} + u_{p,q+1}) = h^2 f_{p,q}$$
$$p=1,\ldots,s,\ q=1,\ldots,r, \qquad (1.1.13)$$

as a finite difference approximation in which u's lying on the boundary are given their prescribed values [20, p.292]. If the equations for the points on the first interior row are written in order, i.e., with q=1, p=1,...,s, followed by the equations for the points on the second, third rows etc., we obtain the matrix-vector form of the above linear system of order rs as

$$Mx = c \qquad (1.1.14)$$

where $M = (M_{ij})_{i,j=1,\ldots,r}$ is in partitioned form having each M_{ij} as a square matrix of order s, defined by

$$M_{ij} = \begin{cases} A & \text{if } i=j, \\ -I_s & \text{if } |i-j|=1, \\ 0 & \text{otherwise} \end{cases}$$

where $A = (a_{ij})$ is a tridiagonal matrix with

$$a_{ij} = \begin{cases} 4 & \text{for } i=j, \\ -1 & \text{for } |i-j|=1, \\ 0 & \text{otherwise} \end{cases}$$

and I_s is the identity matrix of order s (see Bickley and McNamee [173]). It may be observed that x appearing in (1.1.14) has the transpose

$$(u_{11}, \ldots, u_{s1}, u_{12}, \ldots, u_{s2}, \ldots\ldots, u_{1r}, \ldots, u_{sr}).$$

Of course, the elements of c depend on f(x,y) and on the prescribed values of u at the boundary points. It is not difficult to see that (1.1.14) is equivalent to

$$((I_r \otimes A) + (B^T \otimes I_s))x = c \qquad (1.1.15)$$

where $\otimes$ denotes the Kronecker product [4, p.235] and B^T is the transpose of B and $B = (b_{ij})$ is a tridiagonal matrix of order r defined by

$$b_{ij} = \begin{cases} -1 & \text{if } |i-j|=1, \\ 0 & \text{otherwise.} \end{cases}$$

Now (1.1.15) can be converted [251] to the matrix equation AX+XB=C where X is the s x r matrix (u_{ij}) and C is also a matrix of order s x r having its first column as the first s coordinates of c, the second column as the next set of s coordinates of c and so on. Thus the system (1.1.14) involving a large matrix M has been equivalently written as a matrix equation involving matrices of lower dimensions. This suggests that it would be well worth developing algorithms for solving the Sylvester matrix equation rather than its bigger matrix-vector

version of the form (1.1.14) so that storage and computer time can be saved.

We conclude this section with some notational conventions adhered to all through the thesis.

The matrices are denoted by capital letters and, unless the contrary is stated, a matrix is assumed to be square of order n. The (i,j) element of A will be denoted by a_{ij} and we write $A = (a_{ij})$. The identity matrix of order n is denoted by I and if it is of any other order say m, then it is denoted by I_m. The diagonal matrix having d_i as its i-th diagonal entry $(i=1,\ldots,n)$ will be denoted by $\mathrm{diag}(d_1,\ldots,d_n)$. The transpose of a matrix A will be denoted by A' or A^T and the conjugate transpose of A by A^*. We write $\mathrm{Re}(A)$ to denote $(A+A^*)/2$. The trace of A is denoted by $\mathrm{tr}(A)$ and the spectral radius of A by $\rho(A)$.

We denote the class of all $m \times n$ real matrices by $M_{m,n}(\mathbb{R})$ and the class of all $n \times n$ real matrices by $M_n(\mathbb{R})$. The corresponding classes of complex matrices will be denoted simply by $M_{m,n}$ and M_n. $\mathbb{R}^n(\mathbb{C}^n)$ always denotes the n-dimensional real (complex) space of column vectors. We adopt the notation H_n and N_n to represent the sets of all $n \times n$ Hermitian and normal matrices respectively. The symbol Σ_n^+ (Σ_n^-) consistently denotes the class of all positive (negative) definite Hermitian matrices in M_n. With the exception of Sections 3 and 4 of Chapter 3, i.e., Sections 3.3 and 3.4 (in which $A > (\geq) 0$ denotes a matrix

having all positive (nonnegative) entries), elsewhere $P > (\geq) 0$ means that P is a positive definite (semidefinite) Hermitian matrix. To mark the end of a proof of a lemma, theorem or corollary we use ▲▲▲.

1.2. Inertia Theory

The notion of inertia of a matrix was introduced by Ostrowski and Schneider [71] in order to generalize the celebrated result of Lyapunov mentioned in the preceding section and also to study various stability concepts like D-stability, H-stability arising in mathematical economics [1, Chapter 4].

The inertia $In(A)$ of $A \in M_n$ is defined as an ordered triple $(\pi(A), \nu(A), \delta(A))$, the entries denoting the total number of eigenvalues of A with positive, negative, and zero real parts respectively [71]. We have, of course,

$$\pi(A) + \nu(A) + \delta(A) = n. \qquad (1.2.1)$$

In view of the above definition, it is clear that A is stable iff $In(A) = (0,n,0)$. Stable matrices are also termed as negative stable matrices and by a positive stable matrix we mean a matrix whose inertia is $(n,0,0)$.

Theorems involving inertia of a matrix are called the inertia theorems. The purpose of this section is to give a brief account of various inertia theorems. Just recently, Cain [36] has published an excellent survey on inertia theory which also includes some new results. We do not intend to repeat the material presented in that paper. However, we quote

certain important results with an attempt to highlight material which does not feature prominantly in Cain's paper.

The first known inertia theorem is the classical Sylvester's law of inertia. In the following theorem, each statement may be considered as Sylvester's inertia theorem.

THEOREM 1.2.1 (see Cain [36], Carlson [39], Carlson and Schneider [47], Ostrowski and Schneider [71] and Wielandt [88]). The following three statements are true and are equivalent to one another.

(i) If $H \in H_n$ and S is nonsingular, then $In(S^*HS) = In(H)$.

(ii) If $P > 0$ and $H \in H_n$, then $In(PH) = In(H)$.

(iii) If $AH > 0$ and $H \in H_n$, then $In(A) = In(H)$.

Originally Sylvester's theorem appeared in the language of quadratic forms (see Mirsky [18, p.377]). Rado [74] provides a generalization of Sylvester's theorem involving three quadratic forms whereas Schneider [78] gives a topological interpretation of Sylvester's theorem.

Ostrowski [69] obtained a quantitative formulation of Theorem 1.2.1(i) with an extension to singular and rectangular cases in another paper [70]. Thompson [85,86] further generalized these results of Ostrowski. Details of these generalizations will be presented in the forthcoming chapter where we prove some more generalizations of Sylvester's theorem, including a generalization of Thompson's results to normal matrices.

The second known inertia theorem is the classical result of Lyapunov, the one that we have seen in the previous section giving a nice characterization of stable matrices. In this context, we shall provide a list of characterizations of stable matrices in the form of a theorem.

THEOREM 1.2.2. Let $A \in M_n$. Then the following fifteen conditions are equivalent.

(i) A is stable.

(ii) All the solutions of the system $\dot{x} = Ax$ approach zero as $t \to \infty$ [4,6,7,36].

(iii) $\lim_{t\to\infty} e^{At} = 0$.

(iv) For $\alpha > 0$, $A-\alpha I$ is nonsingular and the spectral radius of $(A-\alpha I)^{-1}(A+\alpha I)$ is less than unity [6].

(v) For $\alpha > 0$, $A-\alpha I$ is nonsingular and $(A-\alpha I)^{-1}(A+\alpha I)$ is convergent, i.e., $\{(A-\alpha I)^{-1}(A+\alpha I)\}^k \to 0$ as $k \to \infty$ [20].

(vi) For $\alpha > 0$, $A-\alpha I$ is nonsingular and the sequence $\{x_k\}$ $k=0,1,2,\ldots$ defined by the iterative scheme

$$(A-\alpha I)x_{k+1} = -(A+\alpha I)x_k + 2b$$

converges to the solution of the linear system $Ax = b$ for any initial guess x_0 [20].

(vii) For $C \in \Sigma_n^-$, there exists $H \in \Sigma_n^+$ such that $AH+HA^* = C$ (Lyapunov's theorem) [4,7,13,71,83].

(viii) There exists $H \in \Sigma_n^+$ such that $AH+HA^* \in \Sigma_n^-$ (partly weaker and partly stronger form of (vii)) [10,16,36, 71,83].

(ix) There exists $H \in \Sigma_n^+$ such that $\mathrm{Re}(x^*AHx) < 0$, for all nonzero $x \in \mathbb{C}^n$ [65].

(x) There exists $H \in \Sigma_n^+$ such that the numerical range of AH is contained in the open left half plane [64].

(xi) A is similar to a matrix which can be expressed as the sum of a skew-Hermitian matrix and a negative definite Hermitian matrix [50].

(xii) There exist $H \in \Sigma_n^+$, $Q \in \Sigma_n^-$ and a skew-Hermitian matrix S such that $A = (S+Q)H$ [31].

(xiii) For $C \in \Sigma_n^-$, the matrix equation $A^*H+HA=C$ has a solution and $0 \neq AY+YA^*$ positive semidefinite implies $\mathrm{tr}(Y) < 0$ [32].

(xiv) For $\alpha > 0$, $A-\alpha I$ is nonsingular and for every $C \in \Sigma_n^-$ there exists $H \in \Sigma_n^+$ such that $BHB^*-H = C$ where $B = (A-\alpha I)^{-1}(A+\alpha I)$ (Stein's theorem) [80,83].

(xv) For $\alpha > 0$, $A-\alpha I$ is nonsingular and there exists $H \in \Sigma_n^+$ such that $BHB^*-H \in \Sigma_n^-$ where B is as in (xiv) (partly weaker and partly stronger form of (xiv)) [10,36,80,83].

In literature, there are many generalizations of Lyapunov's theorem. For example, Wong [94] proved that Lyapunov's theorem is valid in the set up of operators. Johnson [64] proved a Lyapunov theorem for angular cones which is a generalization of the formulation (x) of Theorem 1.2.2. Taussky [81] observed some connection between Lyapunov's theorem and D-stability. Also see Reid [76].

Another landmark in the development of inertia theory is the following theorem, what we call the main inertia theorem, proved independently by Ostrowski and Schneider [71] and Taussky [82].

THEOREM 1.2.3. Let $A \in M_n$. Then there exists $H \in H_n$ such that $Re(AH) > 0$ iff $\delta(A)=0$. Moreover, if $H \in H_n$ and $Re(AH) > 0$, then $In(A) = In(H)$.

An alternative proof of this important result is presented by Lancaster [231] using projections.

The interest in the main inertia theorem is that from this the two classical results on inertia due to Sylvester and Lyapunov can be easily deduced [36]. As another corollary of the main inertia theorem, Ostrowski and Schneider [71] proved that if $Re(A) > 0$ and $H \in H_n$, then $In(AH) = In(H)$. This result, known as Wielandt's theorem [88], is in fact equivalent to the second part of the main inertia theorem.

Wimmer [89] has given a shorter proof of the second part of the main inertia theorem, by demonstrating that it is equivalent to the following result.

THEOREM 1.2.4. Let H be a Hermitian matrix partitioned in the form $(H_{ij})_{i,j=1,2}$ where $H_{11} \in \Sigma_m^+$ and $H_{22} \in \Sigma_{n-m}^-$. Then $In(H) = (m,n-m,0)$.

The proof of this theorem essentially depends on Sylvester's theorem. Moreover in proving that Theorem 1.2.4 implies the

second part of the main inertia theorem, Lyapunov's theorem has been used. In view of this it seems to be more appropriate to say that the second part of the main inertia theorem is a consequence of Sylvester's and Lyapunov's theorems. Rather it may be claimed that the second part of the main inertia theorem is equivalent to the two classical results, assuming the fundamental result on the existence and uniqueness of the solution of the Sylvester equation which will be stated in Section 1.4.

It may be noted that Theorem 1.2.4 has appeared as a corollary to the following result proved by Haynsworth [57].

THEOREM 1.2.5. Let H be a Hermitian matrix partitioned in the form $(H_{ij})_{i,j=1,2}$ where H_{11} is nonsingular. Then

$$In(H) = In(H_{11}) + In(K_{22}) \tag{1.2.2}$$

where

$$K_{22} = H_{22} - H_{12}^{*}H_{11}^{-1}H_{12} \tag{1.2.3}$$

and the sum of the inertias is performed by componentwise addition.

Some more results based on this interesting formula may be found in [45,57,58]. In the next chapter, we extend the above theorem to a partitioned normal matrix.

We now turn to the generalizations and applications of the main inertia theorem discussed by various authors. In this connection, we need the concept of controllability of a pair of

matrices. If $A \in M_n$ and $B \in M_{n,m}$, then the pair (A,B) is said to be controllable [90] iff the rank of the $n \times nm$ matrix $(B\ AB\ A^2B\ \ldots\ A^{n-1}B)$ is n. In what follows we use $In(H) \leq In(A)$ in the sense [47] that $\pi(H) \leq \pi(A)$ and $\nu(H) \leq \nu(A)$. This definition is maintained even if H and A are square matrices of different orders.

To start with, Carlson [37,38] and Carlson and Schneider [46,47] studied the main inertia theorem under the situation $Re(AH) \geq 0$ and arrived at the following theorem.

THEOREM 1.2.6. Let $A \in M_n$ and $H \in H_n$. If $\delta(A)=0$ and $Re(AH) \geq 0$, then $In(H) \leq In(A)$. If in addition $\delta(H)=0$, then $In(H) = In(A)$.

Snyders and Zakai [271], Chen [48] and Wimmer [90] have shown that in Lyapunov's theorem and in the main inertia theorem, we may replace "$Re(AH) > 0$" by "$W = Re(AH) \geq 0$ and (A,W) is controllable". Hence we have

THEOREM 1.2.7. If $A \in M_n$ and $H \in H_n$ such that $Re(AH) = W \geq 0$ and (A,W) is controllable, then $\delta(A) = \delta(H) = 0$ and $In(A) = In(H)$.

An extension of this result is given in Wimmer [275]. Using the concept of projection matrices and nilpotent matrices, Wimmer and Ziebur [93] gave a unified treatment of results stated in Theorems 1.2.6 and 1.2.7.

Motivated by Schneider's theorem [77] which is a generalization of Lyapunov's and Stein's theorems, Hill [59,60]

has developed the inertia theory by considering the equation

$$\sum_{i,j=1}^{s} g_{ij}A_iHA_j^* = P \qquad (1.2.4)$$

where $P \in \Sigma_n^+$, $H \in H_n$ and $G = (g_{ij}) \in H_s$ and $A_1,\ldots,A_s$ are quasi-commutative, i.e., each of $A_1,\ldots,A_s$ commutes with $A_iA_j - A_jA_i$ $(i,j=1,\ldots,s)$ and obtained generalizations of many inertia theorems. For $G = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, $A_1 = I$, $A_2 = A$, Hill's theory gives Lyapunov's theorem and the main inertia theorem. For $G = \mathrm{diag}(1,-1)$, $A_1 = I$, $A_2 = A$, it gives Stein's theorem and the Stein analogue of the main inertia theorem. By setting $G = \mathrm{diag}(1,-1,-1,\ldots,-1)$, $A_1 = I$, $A_{i+1} = C_i$, $i=1,\ldots,s-1$, Hill's theory becomes Schneider's theory [77].

By means of a generalized notion of controllability, Carlson and Hill [43] extended Theorem 1.2.7 to Hill's setting given by (1.2.4). Wimmer [92] has considered the specialized form of (1.2.4) with $s=n$ and $A_k = A^{k-1}$, $k=1,\ldots,s$ with $P \geq 0$, to obtain some generalizations of the Lyapunov and Stein theorems. Howland [63] attempted to generalize the main inertia theorem to the equation

$$\sum_{i,j=0}^{n-1} g_{ij}A^iHA^{*j} + \sum_{i,j=0}^{n-1} \bar{g}_{ij}A^jHA^{*i} = P \qquad (1.2.5)$$

where $H \in H_n$ and $P > 0$. Later, Chen [49] disproved Howland's result and obtained some inertia theorems for a less general form than (1.2.5). Pointing out that Chen's results were also false, Carlson and Hill [44] corrected them to prove some generalizations of the main inertia theorem.

The simplified version of Stein's theorem says that for $B \in M_n$, $\rho(B) < 1$ iff there exists $H \in \Sigma_n^+$ such that $BHB^*-H \in \Sigma_n^-$. By applying the Cayley transform $B = (A-I)^{-1}(A+I)$, it was shown by Taussky [83] that Stein's theorem is equivalent to Lyapunov's theorem. By the same technique, the following Stein analogue of the main inertia theorem follows. In order to state the theorem, we define, for $B \in M_n$,

$$In_\Delta(B) = (\pi_\Delta(B), \nu_\Delta(B), \delta_\Delta(B))$$

where the entries in the triplet denote the total number of eigenvalues of B lying respectively outside, inside, and on the unit circle.

THEOREM 1.2.8 (see Cain [36], Hill [60] and Wimmer [89]). Let $B \in M_n$. Then there exists $H \in H_n$ such that $BHB^*-H > 0$ iff $\delta_\Delta(B)=0$. Moreover, if $H \in H_n$ and $BHB^*-H > 0$, then $In_\Delta(B) = In(H)$.

The Stein analogues of Theorems 1.2.6 and 1.2.7 are given in Datta [52] and Wimmer and Ziebur [93] respectively.

A great deal of effort has been devoted by Cain [33-36], to generalize almost all the foregoing results to infinite dimensional setting. These generalizations are extensively discussed in a very recent and expository survey paper on inertia theory by Cain [36].

Hill [61] makes some interesting applications of the main inertia theorem to obtain necessary and sufficient conditions for a matrix to have no eigenvalue on any arbitrary

circle, line and certain other curves in the complex plane. As a corollary to the main inertia theorem, Joyce and Barnett [67] proved that $In(A) = In(H)$ if $A = H^{-1}(S+C)$ where $H \in H_n$, $C > 0$ and S is skew-Hermitian. In the same paper, a sufficient condition on B such that $In(A+B) = In(A)$ has been obtained.

It has been shown by Cain [36] that $In(AH) = In(H)$ for every $H \in H_n$ iff $Re(A) > 0$. In this result, one part is Wielandt's theorem [88] that we have seen earlier and the other part follows immediately from Remark 2 of Carlson's paper [40] on H-stability.

Other studies on inertia of matrices include the following. Inertia theorems for tridiagonal matrices are discussed in Carlson and Datta [41], Datta [51], Schwarz [79] and Wimmer [91] and for Hessenberg matrices in Datta [53].

Johnson [66] has determined the precise set of possible inertias of product of two nonsingular Hermitian matrices with known inertia. Loewy [68] characterized all inertia triples (a,b,c) that are attained by $AH+HA^*$ as H varies over the set of all $n \times n$ positive semidefinite Hermitian matrices. The same study was carried out for positive definite case by Avraham and Loewy [25]. The class of 3×3 real matrices M such that $In(MD) = In(M)$ for all positive diagonal D is characterized by Bahl and Cain [26]. Palmer [72] gives sufficient conditions for an $n \times n$ matrix to have $\delta(A)=0$.

Next, we shall say a few words about the computation of inertia of a matrix. Of course, one way of computing the inertia of A is to find the eigenvalues of A, directly, using an efficient algorithm [22] and then to count the number of eigenvalues having positive, negative, and zero real parts. Without computing the eigenvalues also there are effective algorithms to determine the inertia. It has been shown by Barnett [29] that matrix sign function [54] can be used to compute the inertia of a matrix. If $A = SJS^{-1}$ where J is the quasi-diagonal matrix $\text{diag}(J_+, J_-, J_0)$ with J_+, J_-, J_0 representing the direct sums of the Jordan blocks corresponding to the elementary divisors associated with eigenvalues having positive, negative, and zero real parts respectively [7, Vol.I], then the matrix sign function sgn(A) of A is defined as SDS^{-1} where D is the diagonal matrix having the first $\pi(A)$ diagonal entries as 1, the next $\nu(A)$ diagonal entries as -1, and the last $\delta(A)$ diagonal entries as 0. If A is dichotomic, i.e., if $\delta(A)=0$, then sgn(A) is computed recursively [54,73] by

$$A_{k+1} = (A_k + A_k^{-1})/2, \quad k=0,1,2,\ldots \text{ with } A_0=A. \qquad (1.2.6)$$

Hence it follows that

$$\pi(A) = (n+\text{tr}(E))/2 \text{ and } \nu(A) = (n-\text{tr}(E))/2 \qquad (1.2.7)$$

where

$$E = \text{sgn}(A) = \lim_{k \to \infty} A_k. \qquad (1.2.8)$$

In [42], Carlson and Datta have described an effective computational procedure to compute the inertia of non-Hermitian

matrices. The method is based on the reduction of the given matrix A to a lower Hessenberg matrix with unit codiagonal and then constructing a Hermitian matrix H whose inertia is the same as that of the transformed matrix. An earlier related paper is by Meyer-Spasche [246].

Apart from the results outlined above, there are many other interesting results in inertia theory of matrices. The references given under the title "inertia theory" at the end of the thesis may be consulted for further details.

Finally let us give a brief account of our work related to inertia theory carried out in the present thesis. We introduce a notion of angularity of a matrix (due independently to Cain [36] and Rathore and Chetty [75]) as a generalization of the concept of inertia and prove some angularity theorems concerned with normal matrices. These results include some well-known inertia theorems as special cases. Also we determine the structure of inertia- and angularity-preserving linear transformations on Hermitian, normal and circulant matrices.

1.3. Linear Transformations with Invariants

In recent years, it has been of considerable interest to study the set $L(f,S_n)$ of all linear transformations $T: M_n \to M_n$ such that $T(S_n) \subseteq S_n$ and $f(T(A)) = f(A)$ for all $A \in S_n$ where S_n is some given subset of M_n, for instance, the class of all

Hermitian matrices H_n and f is some given function on M_n, e.g., det(A). In case $T: M_n \to M_n$ has to satisfy only $T(S_n) \subseteq S_n$ then we shall denote the set of all linear transformations T by $L(.,S_n)$.

Within the last two decades a considerable amount of work has been done to determine the structure of these sets L. In most of the cases we see that T has the form

$$T(A) = UAV \text{ for all } A \in M_n \qquad (1.3.1)$$

or

$$T(A) = UA'V \text{ for all } A \in M_n \qquad (1.3.2)$$

for suitable $U, V \in M_n$ where A' denotes the transpose of A.

In this section we present a concise review of the work done till 1980 in the field of linear transformations on matrices. First we mention that there are two excellent expository surveys [117,118] in this area due to Marcus who has made a notable contribution to this sufficiently well-established aspect of linear algebra. These two survey papers may be consulted for further details and references of earlier work in the field.

According to Marcus [118], probably Frobenius initiated the study of determining the structure of $L(f,S_n)$ in 1897, by proving the following result.

THEOREM 1.3.1. Let $T: M_n \to M_n$ be linear such that $\det(T(A)) = \det(A)$. Then T has the form (1.3.1) or (1.3.2) where $\det(UV)=1$.

After a lapse of 28 years, the above result of Frobenius was extended and improved by Schur who considered the linear transformations $T: M_{m,n} \to M_{m,n}$ involving determinants of submatrices. Later, Marcus and May [121] reformulated Schur's result in terms of r-th compound matrix $C_r(A)$ and gave an alternative proof. We recall that if $A \in M_{m,n}$ then $C_r(A)$ is the $\binom{m}{r} \times \binom{n}{r}$ matrix whose elements are the r-th order subdeterminants of A arranged in doubly lexicographic order [16, p.16]. The given alternative proof depends upon the following important result due to Marcus and Moyls [124].

THEOREM 1.3.2. If $T: M_{m,n} \to M_{m,n}$ is linear and maps rank one matrices to rank one matrices then there exist nonsingular matrices U and V in M_m and M_n respectively such that for $m \neq n$

$$T(A) = UAV \text{ for all } A \in M_{m,n} \tag{1.3.3}$$

and for m=n, T has the form (1.3.3) or

$$T(A) = UA'V \text{ for all } A \in M_{m,n}. \tag{1.3.4}$$

The theorem just mentioned was proved using multilinear algebra techniques. Later it has been reproved by Minc [129] using only elementary matrix theory. Also, in the same paper Minc shows that if T preserves the determinant it also preserves rank 1 and thus deduces the classical result of Frobenius, stated in Theorem 1.3.1.

Details of further extensions and applications of rank 1 preservers can be seen in [117,118].

Linear transformations mapping the set of rank k matrices into itself has been studied by Djoković [108]. In [95,96], Beasley has given a variety of sufficient conditions under which the linear map $T: M_{m,n} \to M_{m,n}$ preserving rank k matrices has precisely the same form given in Theorem 1.3.2. To sample, some of the sufficient conditions are

(i) T is nonsingular

(ii) $n \geq k^2 + k$

(iii) $\min(m,n) = k$

(iv) $k = 3$.

A closely related paper about rank k preservers on the space of symmetric matrices is written by Lim [113].

Another problem to which many of the results concerning $L(f,S_n)$ can be reduced is to determine the structure of $L(\text{rank},M_n)$. It has been proved by Marcus and Moyls [125] through a sequence of lemmas that $T \in L(\text{rank},M_n)$ has the form given in Theorem 1.3.2.

A related problem is to find $L(.,S_n)$ with S_n as the class of all nonsingular matrices in M_n. It has been proved by Marcus and Purves [126] that in this case also the conclusion is as in Theorem 1.3.2. Much earlier, Dieudonné [107] concluded the same by taking S_n as the class of all singular matrices in M_n and assuming that T is nonsingular. Recently in 1978, Botta [103] connected the above two results by deriving the result pertaining to nonsingular matrix preservers from the result corresponding to singular matrix preservers.

The structure of $L(\text{rank}, S_n)$ with S_n as the class of all $n \times n$ skew-symmetric matrices has been discussed in Lim [115].

One of the interesting choices of invariant functions in the study of linear preserver problems is the r-th elementary symmetric function $E_r(A)$ of the eigenvalues of $A \in M_n$. It may be noted that $E_1(A) = \text{tr}(A)$ and $E_n(A) = \det(A)$. In this connection, Marcus and Purves [126] established the following in 1959.

THEOREM 1.3.3. If $4 \leq r \leq n-1$, then $T \in L(E_r, M_n)$ has the form (1.3.1) or (1.3.2) where

$$UV = e^{i\phi} I, \quad r\phi \equiv 0 \pmod{2\pi}. \tag{1.3.5}$$

This result says that if $4 \leq r \leq n-1$ and $E_r(T(A)) = E_r(A)$ for all $A \in M_n$, then the linear map $T: M_n \to M_n$ is essentially (modulo taking the transpose and multiplying by a constant) a similarity transformation. This result is not valid for r=1,2 and for this counterexamples have been provided in the same paper. Also for r=n, the above result does not hold. Of course, the latter case corresponds to Theorem 1.3.1. Moreover, this case has been discussed separately by Marcus and Moyls [125] also. The only unsettled case was r=3. In 1970, this was also settled nicely by Beasley [95,97] who proved by an ingenious argument that Theorem 1.3.3 holds for r=3 also.

Kovacs [112] characterized trace-preserving linear maps and also $E_1(A)$-cum-$E_2(A)$-preserving linear maps. The linear transformations on the space of $n \times n$ skew-symmetric real matrices

preserving $E_{2k}(A)$ have been treated by Marcus and Westwick [127].

The problem of determining the structure of $L(h,M_n)$ with the choice of $h(A)$ as the completely symmetric polynomial in eigenvalues of A has been considered in Marcus and Holmes [120] whereas for a large class of polynomials h, the structure of L has been studied by Rackusin and Watkins [140].

By taking $f(A)$ as the trace of the positive semidefinite square root of A^*A (i.e., the sum of the singular values of A), Russo [142] proved that a linear transformation $T: M_n \to M_n$ satisfying $T(I)=I$ and $f(T(A)) = f(A)$ for all $A \in M_n$ has the form (1.3.1) or (1.3.2) with $V = U^*$ and U is unitary.

Defining $\phi_r(A)$ as the r-th elementary symmetric function of the eigenvalues of A^*A, i.e., $E_r(A^*A)$, Marcus and Minc [123] obtained for $1 < r \leq n$ the structure of all linear transformations $T: M_{m,n} \to M_{m,n}$ satisfying $\phi_r(T(A)) = \phi_r(A)$ as the one given in Theorem 1.3.2, with U and V unitary instead of nonsingular.

If $T: M_n \to M_n$ is linear it will be interesting to note that T preserves eigenvalues for all matrices in M_n iff it preserves eigenvalues for all Hermitian matrices in M_n [125]. Moreover in this case T is of the form (1.3.1) or (1.3.2) with UV=I. Consequently, the following theorem was proved by Marcus and Moyls [125]. In the sequel, ev(A) denotes the set of n eigenvalues of A including multiplicities.

THEOREM 1.3.4. Let $T: M_n \to M_n$ be linear. If $T(H_n) \subseteq H_n$ and $ev(T(H)) = ev(H)$ for all $H \in H_n$ then there exists a

unitary $U \in M_n$ such that

$$T(A) = UAU^* \text{ for all } A \in M_n \tag{1.3.6}$$

or

$$T(A) = UA^tU^* \text{ for all } A \in M_n. \tag{1.3.7}$$

We will make use of this result in order to prove one of our main results in Chapter 3.

By characterizing the linear transformation $T: M_n \to M_n$ which leaves both trace and determinant of each matrix A invariant as essentially the similarity transformation of either A or A^t, Minc [128] showed that $T \in L(f,\Delta_n)$, where $f(A)$ is $ev(A)$ and Δ_n is the class of all nonnegative matrices in M_n (i.e., matrices all of whose entries are nonnegative), is of the form

$$T(A) = P^{-1}AP \text{ for all } A \in M_n \tag{1.3.8}$$

or

$$T(A) = P^{-1}A^tP \text{ for all } A \in M_n \tag{1.3.9}$$

where P is a nonnegative generalized permutation matrix. $A \in M_n$ is said to be a generalized permutation matrix if it has precisely one nonzero entry in each row and in each column [128]. In a generalized permutation matrix if all the n nonzero entries are 1, then it is a permutation matrix and if all the n nonzero entries are positive, it is called a nonnegative generalized permutation matrix.

Next, we are concerned with permanent-preservers. If S is the symmetric group of degree n, then the permanent

of $A = (a_{ij}) \in M_n$, denoted by per(A), is defined [122] by

$$\text{per}(A) = \sum_{\sigma \in S} \prod_{i=1}^{n} a_{i\sigma(i)}. \qquad (1.3.10)$$

To begin with, Marcus and May [122] proved that for $n \geq 3$, the linear transformation $T: M_n \to M_n$ such that $\text{per}(T(A)) = \text{per}(A)$ for all $A \in M_n$ has the form

$$T(A) = DPAQL \text{ for all } A \in M_n \qquad (1.3.11)$$

or

$$T(A) = DPA^{t}QL \text{ for all } A \in M_n \qquad (1.3.12)$$

where P, Q are permutation matrices, D, L are diagonal matrices such that per(DL)=1. This result was arrived at after proving nine lemmas. Botta [100] has proved the same result in a somewhat more direct way. Moyls, Marcus and Minc [130] studied the permanent-preservers on the space of doubly stochastic matrices whereas Lim and Ong [114] studied the same on the space of real symmetric matrices. In [136], Pierce considered discriminant-preserving linear maps.

The structure of $L(f, M_n)$ with f as the generalized matrix function in the sense of Schur, i.e.,

$$f(A) = \sum_{\sigma \in G} \lambda(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)}, \; A = (a_{ij}) \in M_n \qquad (1.3.13)$$

where λ is a nonzero function defined on a subgroup G of the symmetric group acting on $\{1, \ldots, n\}$, has been investigated in Botta [101,102] and Ong [132]. Ong and Botta [133] and Ong [131] studied linear maps preserving the class of generalized permutation matrices.

In 1976, Watkins [144] proved that for $n \geq 4$, any nonsingular linear map $T: M_n \to M_n$ satisfying "$AB=BA$ implies $T(A)T(B)=T(B)T(A)$ for all A and B in M_n" (i.e., T preserving commuting pairs of matrices) has the form

$$T(A) = cU^{-1}AU + g(A)I \text{ for all } A \in M_n \tag{1.3.14}$$

or

$$T(A) = cU^{-1}A'U + g(A)I \text{ for all } A \in M_n \tag{1.3.15}$$

for some scalar c, nonsingular matrix U and linear functional g. The invalidity of this result for $n=2$ was shown by means of counterexample. Once again it was Beasley [98] who settled the problem by showing that the result of Watkins holds for $n=3$.

The k-th numerical range of $A \in M_n$ denoted by $W_k(A)$ is defined by

$$W_k(A) = \{\sum_{i=1}^{k}(Av_i,v_i)\} \tag{1.3.16}$$

where $\{v_1,\ldots,v_k\}$ runs through all orthonormal sets. The k-th decomposable numerical range of A is defined to be the set

$$\hat{W}_k(A) = \{\det(X^*AX): X \in M_{n,k} \text{ and } \det(X^*X)=1\}. \tag{1.3.17}$$

It may be observed that in both cases $k=1$ gives the classical numerical range. Pierce and Watkins [137] considered the problem of determining all linear transformations $T: M_n \to M_n$ which preserve $W_k(A)$. Just recently, Marcus and Filippenko [119] treated the corresponding problem for $\hat{W}_k(A)$ and proved that T is of the form

$$T(A) = \xi U^*AU \text{ for all } A \in M_n \qquad (1.3.18)$$

or

$$T(A) = \xi U^*A'U \text{ for all } A \in M_n \qquad (1.3.19)$$

where ξ is a complex k-th root of unity and U is unitary. This result is a generalization of a result of Pellegrini [134] who characterized the linear operators preserving the classical numerical range.

Next we come to unitary-preserving maps. It is proved by Marcus [116] that $T \in L(.,U_n)$, where U_n is the class of all $n \times n$ unitary matrices, is of the form (1.3.1) or (1.3.2) where U and $V \in U_n$. Botta [104] gives a new proof of this result under the special assumptions that T is nonsingular and $n \geq 3$. It was conjectured by Marcus [118, Conjecture 6] that if $T: M_n(\mathbb{R}) \to M_n(\mathbb{R})$ is linear such that T maps the real orthogonal group G into itself then T must have the form (1.3.1) or (1.3.2) in which U, V belong to G. Wei [145] showed that this conjecture holds except for n = 2, 4, or, 8 and that in the exceptional cases there exist singular maps. Linear operators preserving certain algebraic groups are discussed in Pierce [135] and Botta and Pierce [105].

A linear transformation $T: M_n \to M_n$ is said to be Hermitian-preserving iff $T(H_n) \subseteq H_n$. de Pillis [138] shows that Hermitian-preserving linear transformations T have the structure given by

$$T: A \to \Sigma\, \alpha_i X_i^* A' X_i$$

where each α_i is real and X_i is a certain matrix in M_n depending on T. Hill [110] gives three sets of necessary and sufficient conditions for T to be Hermitian-preserving. Some more papers related to Hermitian-preserving maps are Choi [106] and Poluikis and Hill [139].

It is shown by Schneider [77] that a linear transformation T on the real space $\tilde{H}_n$ of n x n Hermitian matrices (i.e., a space in which a matrix may be multiplied only by a real scalar) taking the cone of positive semidefinite matrices onto itself preserves rank. It is further shown that in this case there exists a nonsingular C such that

$$T(H) = CHC^* \text{ for all } H \in \tilde{H}_n \qquad (1.3.20)$$

or

$$T(H) = CH'C^* \text{ for all } H \in \tilde{H}_n. \qquad (1.3.21)$$

If T is a linear transformation on the space of n x n real symmetric matrices S_n into itself sending real positive definite matrices into themselves and satisfying $\det(T(A)) = c(\det(A))$ where c is a nonzero real constant, then Eaton [109] proves that there exists a real nonsingular matrix M such that

$$T(A) = MAM' \text{ for all } A \in S_n. \qquad (1.3.22)$$

Sinkhorn [143] and Benson [99] characterized $T \in L(., \Omega_n)$ where Ω_n represents the class of all generalized doubly stochastic matrices in M_n (i.e., n x n complex matrices whose row and column sums are 1) as linear combinations of functions of the type

$$T(X) = AXB \qquad (1.3.23)$$

with restrictions posed on the n x n matrices A and B.

It has been shown by Beasley [95] that linear maps $T: M_n \to M_n$ preserving minimal polynomial has the form (1.3.1) or (1.3.2) with UV = I. The structure of linear maps that preserve matrices annihilated by a polynomial is the study of Howard [111].

Recently, Robinson [141] has shown that results characterizing the structure of $L(\text{rank},M_n)$, $L(\det,M_n)$ and $L(\text{ev},M_n)$ can have direct generalizations to multilinear functions on M_n. Specifically, if f(X) is the rank, determinant or the set of eigenvalues including multiplicities of X and m is a positive integer, the set of m-linear functions T from the Cartesian product $M_n \times M_n \times \ldots \times M_n$ (m copies) to M_n satisfying

$$f(T(X_1,\ldots,X_m)) = f(X_1 \ldots X_m) \text{ for all } X_1,\ldots,X_m \in M_n \quad (1.3.24)$$

has been determined.

To get an overall picture about the structure of $L(f,S_n)$ or $L(.,S_n)$ for various f and S_n, we shall tabulate below the results presented in the survey for which $T: M_n \to M_n$ is of the form (1.3.1) or (1.3.2). In the table, we use the following notation. R_1 is the set of all rank 1 matrices, GL_n is the set of all nonsingular matrices, SL_n is the set of all singular matrices, U_n is the set of all unitary matrices, Δ_n is the set of all nonnegative matrices, P_n is the set of all permutation matrices, NGP_n is the set of all nonnegative generalized permutation matrices and D_n is the set of all diagonal matrices. All these sets are subsets in M_n.

TABLE 1.3.1

Structures of linear transformations on matrices with some invariants

S.No.	f(A)	S_n	Additional assumption	Condition on U and V in (1.3.1) and (1.3.2)
1.	det(A)	M_n	-	det(UV) = 1
2.	-	R_1	-	$U,V \in GL_n$
3.	rank(A)	M_n	-	$U,V \in GL_n$
4.	-	GL_n	-	$U,V \in GL_n$
5.	-	SL_n	T is nonsingular	$U,V \in GL_n$
6.	$E_r(A)$	M_n	$3 \leq r \leq n-1$	$UV = e^{i\phi}I$, $r\phi \equiv 0 \pmod{2\pi}$
7.	$tr\{(A^*A)^{1/2}\}$	M_n	$T(I) = I$	$UV = I$, $U \in U_n$
8.	$E_r(A^*A)$	M_n	$1 < r \leq n$	$U,V \in GL_n$
9.	ev(A)	M_n	-	UV = I
10.	ev(A)	H_n	-	$UV = I$, $U \in U_n$
11.	trace-cum-determinant	M_n	-	UV = I
12.	ev(A)	Δ_n	-	$UV = I$, $U \in NGP_n$
13.	per(A)	M_n	$n \geq 3$	$U = DP$, $V = QL$ where $P,Q \in P_n$ $D,L \in D_n$ such that per(DL) = 1
14.	-	U_n	-	$U,V \in U_n$
15.	minimal polynomial of A	M_n	-	UV = I

We may also mention here that in Chapter 3 of the present thesis we determine the classes $L(In,H_n)$, $L(\theta,N_n)$, $L(In,N_n)$, $L(\theta,C_n)$ and $L(In,C_n)$ where In and θ stand for inertia and angularity respectively and H_n, N_n, C_n denote respectively the classes of all Hermitian, normal and circulant matrices in M_n. Also we determine the structure of linear transformations on $\mathbb{R}^n$ and $\mathbb{C}^n$ preserving certain qualitative and quantitative invariants.

1.4. The Matrix Equation AX+XB=C

Our concern in this section is with the linear matrix equation

$$AX + XB = C \qquad (1.4.1)$$

where A, B, C are known complex matrices of order $m \times m$, $n \times n$ and $m \times n$ respectively. It may be observed that A and B should necessarily be square matrices for the above equation to be conformable. In literature, the equation (1.4.1) is referred as the Sylvester equation and for the choice $B=A^*$, it is well known as the Lyapunov matrix equation.

During the past quarter of a century, these two equations, with more emphasis on real type, have received a great deal of attention because of their practical importance in a variety of problems, especially in control theory. The solutions of these types of equations are required in

(i) solving by finite difference discretization, some boundary value problems in partial differential equations which occur, for example, in potential

theory or when finding the stress in a helical spring [173,211,212],

(ii) the analysis of beam gridworks with various boundary conditions [239],

(iii) the study of certain type of linear ordinary differential system with constant coefficients [184],

(iv) the investigation of the stability of time-invariant systems [170,223],

(v) the calculation of quadratic performance indices for linear time-invariant systems [169,223],

(vi) the calculation of mean-square functionals of linear time-invariant systems [224],

(vii) the calculation of a large class of functionals of the time and frequency response of a linear, constant coefficient dynamical system [240],

(viii) estimating shoots and settling times and deriving pseudo-optimal control policies for the vector u which will return the system with control to equilibrium as quickly as possible following an initial disturbance [254],

(ix) the construction of Luenberger observers for linear time-invariant multivariable systems [236],

(x) sensitivity analysis of optimal linear control systems to small variations in parameters [153],

(xi) the evaluation of ISE (integral of the square error) of certain control systems [178],

(xii) simplification of large dynamical systems [178], and

(xiii) the computation of inertia of certain type of matrices [42,246].

For some more applications one may refer Barnett and Storey [160], Chidambara and Viswanadham [178] and Rothschild and Jameson [264].

The Sylvester equation and its simpler more structured cousin, namely, the Lyapunov equation have been thoroughly studied and the literature in this area is quite extensive. Our objective is to give a comprehensive account of existence theorems, various methods of solutions, different types of solutions including explicit solutions and numerical solutions, sensitivity of solution, comparative studies made on several algorithms and some generalized forms of the matrix equation AX+XB=C. A guide to our survey in this area is the excellent treatment of Lancaster [231] in this direction. Other papers worth noting in this connection are those of Kučera [228] who has presented a comprehensive theory of AX+XB=C and of Barnett and Storey [160] concerning various solution methods for the Lyapunovmatrix equation.

We shall begin our survey with various existence theorems. The straightforward approach to solve (1.4.1) is to transform it to an equivalent vector form. To facilitate this study, we need the concepts of column string of a matrix [273] and the Kronecker product of two matrices (see, e.g., Bellman [4, p.235] and Lancaster [13, p.256]).

If $X = (x_{ij}) \in M_{m,n}$, then the column string of X, written cs(X) is defined as the mn-dimensional column vector

$$(x_{11},\ldots,x_{m1},x_{12},\ldots,x_{m2},\ldots\ldots,x_{1n},\ldots x_{mn})^T.$$

Neudecker [251] refers the column string of X as the vector of X denoted by vec(X).

If $A = (a_{ij}) \in M_{m,n}$ and $B = (b_{ij}) \in M_{p,q}$ then the Kronecker product (also known as tensor product or direct product) of A and B, denoted by $A \otimes B$, is defined to be the partitioned matrix $(a_{ij}B) \in M_{mp,nq}$ $(i=1,\ldots,m,\ j=1,\ldots,n)$.

Some elementary but interesting properties of the Kronecker product follow. In (1) and (3), A and B may be rectangular matrices of any order whereas in the remaining it is assumed that $A \in M_m$, $B \in M_n$.

(1) $(A \otimes B)^T = A^T \otimes B^T$

(2) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, if A and B are nonsingular

(3) $\text{rank}(A \otimes B) = \text{rank}(A)\,\text{rank}(B)$

(4) $\text{tr}(A \otimes B) = \text{tr}(A)\,\text{tr}(B)$

(5) $\det(A \otimes B) = (\det(A))^n(\det(B))^m$

The following important classical result on the eigenvalues of the Kronecker products is given in Lancaster [13, p.259].

Let $\phi(x,y) = \sum_{i,j=0}^{p} c_{ij}x^iy^j$ be a polynomial in x and y for certain complex numbers c_{ij}. Then for $A \in M_m$ and $B \in M_n$, the Kronecker polynomial $\phi(A,B)$ is defined as $\sum_{i,j=0}^{p} c_{ij}A^i \otimes B^j$ where as usual $A^0 = I_m$ and $B^0 = I_n$. Then if $\alpha_1,\ldots,\alpha_m$ are

the eigenvalues of A and $\beta_1,\ldots,\beta_n$ are those of B, then the eigenvalues of $\phi(A,B)$ are the mn numbers $\phi(\alpha_i,\beta_j)$, i=1,...,m, j=1,...,n. In particular the eigenvalues of $A \otimes B$ are $\alpha_i\beta_j$ i=1,...,m, j=1,...,n and consequently $\rho(A \otimes B) = \rho(A)\rho(B)$.

The following result connecting the column string and the Kronecker product is given in Vetter [273] (also see Marcus and Minc [16, p.9]). If A, B, X are matrices of any order so that AXB is defined then

$$cs(AXB) = (B^T \otimes A)\, cs(X). \tag{1.4.2}$$

Hence, by making use of this formula, we find that the Sylvester equation (1.4.1) assumes an equivalent form

$$Gx = c \tag{1.4.3}$$

where

$$G = (I_n \otimes A) + (B^T \otimes I_m),\ x = cs(X),\ c = cs(C).$$

Since the eigenvalues of G are $\alpha_i+\beta_j$, i=1,...,m, j=1,...,n where $\alpha_1,\ldots,\alpha_m$ are the eigenvalues of A and $\beta_1,\ldots,\beta_n$ are those of B, one has the following basic existence and uniqueness theorem about the Sylvester equation.

THEOREM 1.4.1. The solution of the Sylvester equation AX+XB=C exists and is unique iff A and -B do not have a common eigenvalue.

Alternative proofs for this important theorem are available in the classical texts of Gantmacher [7, Vol.I] and MacDuffee [15]. For a simpler and interesting proof, see Ostrowski and Schneider [71].

In view of (1.4.3), we conclude that (1.4.1) has a solution iff

$$\text{rank}(G \mid \text{cs}(C)) = \text{rank}(G). \tag{1.4.4}$$

Analogous existence theorem is given by Roth [262] even for the more general linear matrix equation

$$\sum_i A_i X_i B_i = C \tag{1.4.5}$$

and in particular for

$$\sum_i A_i X B_i = C. \tag{1.4.6}$$

One may feel the difficulty in expressing the condition given in (1.4.4) in terms of A, B, C explicitly. Moreover, the matrix G involved here is of order $mn \times mn$. In another paper, involving only matrices of order $(m+n)$, Roth [263] has given a necessary and sufficient condition for (1.4.1) to have a solution. The result is as follows.

THEOREM 1.4.2. Let $A \in M_m$, $B \in M_n$ and $C \in M_{m,n}$. Then there exists $X \in M_{m,n}$ such that $AX+XB=C$ iff the matrices $\begin{bmatrix} A & C \\ 0 & -B \end{bmatrix}$ $\begin{bmatrix} A & 0 \\ 0 & -B \end{bmatrix}$ are similar.

In fact, Roth has proved this theorem only for the case $m=n$, though this holds in general. Another result associated with the above theorem says that for $A \in M_{m,r}$, $B \in M_{s,n}$ and $C \in M_{m,n}$, there exist $X \in M_{r,n}$ and $Y \in M_{m,s}$ such that $AX+YB=C$ iff $\begin{bmatrix} A & C \\ 0 & -B \end{bmatrix}$ and $\begin{bmatrix} A & 0 \\ 0 & -B \end{bmatrix}$ are equivalent in the sense that one can be obtained from the other by means of a sequence of elementary row and column operations.

The above two results, known as Roth's similarity theorem and Roth's equivalence theorem, have been discussed by several authors [148,187,188,195,196,201,218,228] either by giving alternative proof or by extending these results over certain rings.

Johnson and Newman [216] proved that if A and B are diagonalizable then AX+XB=C has a solution iff $\begin{bmatrix} A & C \\ 0 & -B \end{bmatrix}$ is diagonalizable. It may be noted that this result is an immediate corollary of Roth's similarity theorem.

In [228], Kučera obtained two existence theorems regarding the equation (1.4.1), one involving linear transformations and the other one being Roth's similarity theorem. Criterion for the existence of a solution of the Sylvester equation is given in terms of pseudo-inverse of matrices by Boullion and Poole [175]. Some more remarks on existence and uniqueness theorems are given in Lancaster [231].

Next we shall deal with various methods for solving the Sylvester equation. The basic method is to solve the transformed equation (1.4.3) by any standard method (see e.g., Westlake [21] and Young [23]). This direct approach known as full inversion method [264] is satisfactory only for small values of m and n, say less than 10. Since the computer time and storage requirements of this method increase with $(mn)^3$ and $(mn)^2$ respectively, this method is of no value even for modest values of m and n. Thus in order to solve AX+XB=C

many authors have suggested special methods which involve operations only with matrices of order m and n.

The best known such methods are based on the transformation of A and B to simple forms (for example, companion form, Jordan form, Hessenberg form, Schwarz form [223] etc.) via similarity transformations

$$\tilde{A} = U^{-1}AU, \quad \tilde{B} = V^{-1}BV. \tag{1.4.7}$$

Now (1.4.1) reduces to

$$\tilde{A}Y + Y\tilde{B} = \tilde{C} \tag{1.4.8}$$

where $Y = U^{-1}XV$ and $\tilde{C} = U^{-1}CV$. In view of the special structures of $\tilde{A}$ and $\tilde{B}$, solving (1.4.8) may be shown to be simple and therefrom X can be obtained easily.

The well-established algorithm of Bartels and Stewart [164] and its recent improvement by Golub, Nash and Van Loan [192] are examples of the above method. Key to the technique of the Bartels-Stewart algorithm is the Schur reduction to triangular form by orthogonal similarity transformations using standard procedures of Householder's method and the QR algorithm. More specifically, $\tilde{A}$ is in block lower (upper) triangular form and $\tilde{B}$ is in block upper (lower) triangular form where in both cases each block is of order at most two. Now the structure of transformed equation (1.4.8) allows the solution process to be decoupled and reduced to a succession of the Sylvester equation problems with matrices of order 2×2 at most. Then each such equation can be solved by direct approach, that is

by full inversion method explained earlier. A special feature of the Bartels-Stewart algorithm is that it contains a package of FORTRAN IV program with brief description of the subroutines.

The above algorithm requires approximately $(2+4\sigma)(m^3+n^3) + \frac{5}{2}(mn^2+nm^2)$ multiplications where σ is the number of QR steps involved in the process. The storage requirement for this method is $2m^2+2n^2+mn$ locations.

Since the computer time required mainly depends on the number of multiplicative operations, only multiplications are considered in the operation count. Moreover all the counts are made on the assumption that the matrices concerned are all real.

Recently, Golub et al. [192] and Enright [185] proposed independently a modification over the Bartels-Stewart algorithm by keeping A in Hessenberg form itself. The rest of the procedure is as in the Bartels-Stewart algorithm. Assuming that the Schur reduction of B requires $10n^3$ operations, the work count for the new method, called the Hessenberg-Schur algorithm [192], has been calculated as $\frac{5}{3}m^3+10n^3+5m^2n+\frac{5}{2}mn^2$. Although the storage requirements of this modified method exceeds that of the Bartels-Stewart algorithm by m^2 locations, it has been shown in [192] that it is considerably faster than its nearest competitor, the Bartels-Stewart algorithm. The stability of the new method is demonstrated through a roundoff error analysis and supported by numerical tests.

The transformation approach of solving the Sylvester equation has been discussed by several authors. Gantmacher [7, Vol.I, Chapter VIII] uses this approach to study the general solution of (1.4.1). Molinari [249] considers $\tilde{A}$ and $\tilde{B}$ in companion form and estimates the number of operations required as about $2(m+n)(m^2+n^2)$ and storage needed as $6m^2+6m$ locations. Though this procedure is countwise favourable, there is a possibility of obtaining ill-conditioned transformation matrices [192]. The approach of Guidorzi [194] deals with general canonical forms and it is suggested that the transformed equation (1.4.8) may be solved using direct approach by triangularizing the almost triangular matrix of order mn. Kreisselmeier's method [227] involves a transformation of B (or A^T) to Hessenberg form, an m x m (or n x n) matrix inversion and a recursive algorithm. Treating $\tilde{A}$ and $\tilde{B}$ as Jordan canonical forms, Ma [239] obtains a solution involving a finite double matrix series.

The transformation technique can also be used to study certain explicit solutions of AX+XB=C. In (1.4.8), if $Y = (y_{ij})$, $\tilde{C} = (\tilde{c}_{ij})$, $\tilde{A} = \text{diag}(\alpha_1,\ldots,\alpha_m)$ and $\tilde{B} = \text{diag}(\beta_1,\ldots,\beta_n)$ then it follows that

$$y_{ij} = \tilde{c}_{ij}/(\alpha_i+\beta_j) \qquad (1.4.9)$$

provided $\alpha_i+\beta_j \neq 0$. This shows that if A and B are diagonalizable and A and -B do not have a common eigenvalue then the unique solution of (1.4.1) can be expressed in terms of eigenvalues and eigenvectors of A and B [173,215].

It is, however, possible to obtain an explicit solution without recourse to diagonalization. It has been established by Bickley and McNamee [173] that if the characteristic polynomial of B is $x^n - p_1 x^{n-1} + \ldots + (-1)^n p_n$, then the solution to the Sylvester equation is given explicitly as

$$X = G^{-1} \sum_{r=1}^{n} (-1)^{r-1} A^{n-r} C B_{(r-1)} \qquad (1.4.10)$$

where

$$G = A^n + p_1 A^{n-1} + \ldots + p_n I_m \qquad (1.4.11)$$

and

$$B_{(r)} = B^r - p_1 B^{r-1} + \ldots + (-1)^r p_r I_n, \quad r=0,1,\ldots,n. \qquad (1.4.12)$$

Similarly X can be expressed depending on the characteristic polynomial of A. Explicit solutions of this nature involving the coefficients of characteristic polynomial of A (or B) and an inversion of a matrix of order n (or m) are found with slight variations in Barnett [154], Jameson [215], Lu [234] and Šestopal [266]. Bickart [172] obtains an explicit solution involving coefficients of annihilating polynomials of A or B. Expressions for the solution of (1.4.1) involving the matrix $\begin{bmatrix} A & C \\ 0 & -B \end{bmatrix}$ and the characteristic polynomials of A and B are available in Jones [220,221]. Müller [250] uses the resultant of the characteristic polynomials of A and -B to express the solution of (1.4.1). Hartwig [199] derives a finite series solution for X in case C can be expressed as a finite series in terms of matrices related to A and B. In another paper [200], Hartwig applies the theory of generalized

inverses to obtain explicit solutions for the Sylvester equation in case A and -B have a common eigenvalue. This singular case is studied in Ma [239] also.

For details of solutions in terms of component matrices [13, p.174] and solutions in terms of adjoint matrices refer Lancaster [231].

The equation AX+XB=C can be expressed [270] in the equivalent form

$$X - UXV = W \tag{1.4.13}$$

with

$$U = (qI_m - A)^{-1}(qI_m + A) \tag{1.4.14}$$

$$V = (qI_n + B)(qI_n - B)^{-1} \tag{1.4.15}$$

and

$$W = -2q(qI_m - A)^{-1}C(qI_n - B)^{-1} \tag{1.4.16}$$

for a suitable nonzero constant q. The form (1.4.13) suggests an infinite series solution to (1.4.1) given by

$$X = \sum_{j=1}^{\infty} U^{j-1}WV^{j-1}. \tag{1.4.17}$$

In fact, the above series converges [231] iff $\rho(U)\rho(V) < 1$ and in this case the limit represents the unique solution of the Sylvester equation. Based on the representation (1.4.17), Smith [270] proposes the quadratic convergence iterative scheme

$$X_{k+1} = X_k + U^{2^k} X_k V^{2^k}, \quad X_o = W \tag{1.4.18}$$

to obtain the numerical solution of (1.4.1), assuming that

A and B are stable. Under this assumption, $\rho(U) < 1$ and $\rho(V) < 1$ for any $q > 0$.

Another interesting and significant form of explicit solution is the one expressed in terms of integrals. Suppose Γ_A and Γ_B are contours enclosing in their interiors respectively all α_i and β_j such that $\alpha_i+\beta_j \neq 0$ and let $P_A C P_B = C$ where $\{\alpha_i\}$ and $\{\beta_j\}$ are the eigenvalues of A and B respectively and

$$P_A = \frac{1}{2\pi i} \int_{\Gamma_A} (\alpha I_m - A)^{-1} d\alpha \qquad (1.4.19)$$

$$P_B = \frac{1}{2\pi i} \int_{\Gamma_B} (\beta I_n - B)^{-1} d\beta . \qquad (1.4.20)$$

Then

$$X = -\frac{1}{4\pi^2} \int_{\Gamma_A} \int_{\Gamma_B} \frac{(\alpha I_m - A)^{-1} C (\beta I_n - B)^{-1}}{\alpha+\beta} \, d\beta \, d\alpha \qquad (1.4.21)$$

is a solution of (1.4.1). Moreover, if $\mathrm{Re}(\alpha_i+\beta_j) < 0$ for all the enclosed eigenvalues, then (1.4.21) can be expressed as

$$X = -\int_0^\infty e^{At} C e^{Bt} \, dt. \qquad (1.4.22)$$

If $\alpha_i+\beta_j \neq 0$ for all i and j, i.e., if $\sigma(A) \cap \sigma(-B) = \phi$, the empty set, where $\sigma(A)$ denotes the spectrum of A, then (1.4.21) represents the unique solution of (1.4.1) and in this case $P_A = P_B = I$ and hence the requirement $P_A C P_B = C$ is automatically satisfied. If $\mathrm{Re}(\alpha+\beta) < 0$ for all $\alpha \in \sigma(A)$ and $\beta \in \sigma(B)$ (which holds in particular if A and B are stable), then (1.4.22) is the unique solution of (1.4.1). These results involving integrals are due to Krein [12] and they

are quoted in Kučera [228] and discussed extensively in Lancaster [231].

Using differential equations, Bellman [4, p.179] proves that if the integral given in (1.4.22) exists for all C, then it represents the unique solution of AX+XB=C. Many authors use the formula (1.4.22) to obtain numerical solutions of the Sylvester and Lyapunov equations. Regarding this, references will be given later.

A number of other papers dealing with the methods of solution of AX+XB=C may be recognized from their titles. Powers [259] proposes a method using two ideas familiar to practitioners of control theory: controllability of a matrix and 'tearing'. This method seems to be not so favourable since it involves $O(n^4)$ operations, for m=n. Bar-Ness and Langholz [152] investigates the solution of (1.4.1) as an eigenvalue problem of $\begin{bmatrix} B & O \\ C & -A \end{bmatrix}$. Ingraham and Trimble [214] reduce the problem of solving (1.4.1) to a problem in polynomial congruences.

There may be situations in which we require certain specific type of solution to AX+XB=C. For example, finding nonsingular solution arises directly in the construction of an observer for linear time-invariant system [236]. Hearon [202] gives a necessary and sufficient condition for the Sylvester equation to possess nonsingular solution. In view of assignment problems, Boullion and Poole [175] investigate the

existence of integral solution whereas Kabe [222] considers the nonnegative and integral solutions.

In [271], Snyders and Zakai discuss nonnegative definite solutions of (1.4.1) with $B = A^*$ and $C = -DD^*$. If $AX+XB=C$ is not consistent, the reasonable compromise is to find a least square solution: that is to find X which minimizes the Frobenius norm of the residual matrix $AX+XB-C$ and this problem has been treated in Lovass-Nagy and Powers [233].

For solving (1.4.1), Milani [247] presents an iterative procedure which explores possibilities of partitioning the coefficient matrices A, B and C for decomposition of the equation into sufficiently lower dimension equations which are satisfactorily solved by direct method. Varah [272] investigates the feasibility of the iterative scheme

$$AX_{k+1} = -X_kB + C \qquad (1.4.23)$$

for solving (1.4.1). It is easily shown that the above method converges for any initial guesss X_0 iff $\rho(A^{-1})\rho(B) < 1$. The sensitivity aspects of the solution of (1.4.1) have been studied in Golub, Nash and Van Loan [192] and Varah [272].

Next, we shall make a brief mention of studies on some generalized forms and special cases of the Sylvester equation. The general matrix equation $\sum_i A_iXB_i = C$ has already been referred in connection with existence theorems. The history of this general matrix equation is given in MacDuffee [15]. The discussion made by Lancaster [231] on this equation is

supplemented by Wimmer and Ziebur [276]. Vetter [273] deals with the yet more general equation

$$\sum_i A_i X B_i + \sum_i C_i X^T D_i = F. \qquad (1.4.24)$$

In particular, many authors have concentrated on the matrix equation

$$AXB + CXD = E. \qquad (1.4.25)$$

The above equation occurs in the MINQUE theory of estimating covariance components in a covariance components model [248]. The general two-layer gridwork problems with different boundary conditions all lead to the matrix equation of the form (1.4.25) [238]. Moreover, the solution of this type of equation is required in linear parametric estimation theory of normal multivariate statistical analysis [265] and in the numerical solution of certain implicit ordinary differential equations [186].

Mitra [248] describes a method of solution to (1.4.25) using canonical representation of a singular pencil studied in Gantmacher [8]. Baksalary and Kala [149] give a necessary and sufficient condition for (1.4.25) to be consistent, together with a representation of its general solution in terms of g-inverses for a consistent case. Epton [186] extends the idea of Enright [185], referred earlier, to solve (1.4.25). The equation (1.4.25) is considered, also in Golub et al. [192], Jones [219] and Scobey and Kabe [265].

Another generalization of the Sylvester equation is

the matrix Riccati equation (or matrix quadratic equation)

$$XDX + AX + XB - C = 0. \qquad (1.4.26)$$

Interest in this type of equation stems again from its wide application in electrical engineering problems [146,255]. This equation has been studied in Anderson [146], Beavers and Denman [165,166], Campbell and Daughtry [176], Coppel [179], Daughtry [180], Jones [217], Jones [220], Kleinman [226], Laub [232], Mårtensson [242], Meyer [243,244], Potter [255] and Wimmer [275]. For additional references on this equation, one may consult the very good source book of Barnett [1] on matrices in control theory.

In view of the applications in quantum mechanics, scattering theory and in the study of similarity of operators [190], a great deal of work has been done on the operator equations corresponding to (1.4.1) and (1.4.6) by Apostol [147], Freeman [190], Goldstein [191], Luenberger [235], Lumer and Rosenblum [237] and Rosenblum [260,261].

Several authors, for example, Foulkes [189], Gantmacher [7, Vol.I], Hartwig [198], MacDuffee [15] and Parker [253] have studied the equation (1.4.1) with C=0. In addition, if B=-A then the problem of solving (1.4.1) reduces to find all matrices X commuting with A. This problem dates back to Frobenius and has been solved by many authors [228]. For some references on this study see Bellman [4, p.30]. Another special case of (1.4.1) is AX=C.

The most important special case of the Sylvester equation is the one mentioned in the beginning of this section. It is the Lyapunov matrix equation

$$AX + XA^* = C, \tag{1.4.27}$$

which is also known as the continuous Lyapunov matrix equation, it being associated with continuous-time linear system $\dot{x} = Ax$. There have been many papers on the Lyapunov matrix equation, especially on the numerical solution of the real matrix equation

$$AX + XA^T = C \tag{1.4.28}$$

where $A, C \in M_n(\mathbb{R})$. We have seen earlier that the solution of (1.4.28) with $C = -I$ can effectively be used to determine the stability of a real matrix A. Here the essential problem is to solve

$$AX + XA^T = - I \tag{1.4.29}$$

admitting that there is a symmetric solution X [223]. Now the system (1.4.29) represents $n(n+1)/2$ linear equations for the $n(n+1)/2$ unknown elements in X. The systematic way of constructing the enlarged system of order $n(n+1)/2$ is given in Bingulac [174], Chen and Shieh [177] and MacFarlane [240]. Barnett and Storey [157] have shown that the number of equations can be reduced to $n(n-1)/2$ by introducing a skew-symmetric matrix. It has been shown by calculations [158] that this reduction in the number of equations is worthwhile for moderate values of n, say lying between 10 and 50.

Now we shall investigate various methods of solutions to the Lyapunov matrix equation. Smith [269] derives an explicit expression for X satisfying (1.4.27) and Ziedan [278] develops an explicit solution for (1.4.28) involving Schwarz form of A.

Since the direct approach suffers most with respect to increase of computer time, the transform method, for instance, the Bartels-Stewart algorithm, is preferable. Power [256] describes an iterative method for solving (1.4.28) when A is given in Schwarz or Routh canonical form. Howland and Senez [213] have described a method for solving (1.4.29) when A is in upper Hessenberg form and this procedure has been extended to complex case by Meyer-Spasche [245].

Although the solution of the Lyapunov matrix equation is used to solve the stability problem, in many applications, the solution of (1.4.28) is required only for the case when A is stable. In such a situation there are many efficient algorithms. Barnett and Storey [159] remark that the most promising method from the practical point of view is the one based on the iterative scheme given in (1.4.18) suggested by Smith [270]. This method has been discussed by another Smith [267] also.

In the light of numerical quadrature, Davison and Man [183] have proposed the following iterative procedure to solve (1.4.28) when A is stable:

$$X_{k+1} = Q^{2^k} X_k (Q^T)^{2^k} + X_k \ , \ X_o = -hC \qquad (1.4.30)$$

where

$$Q = (I - \frac{h}{2}A + \frac{h^2}{12}A^2)^{-1}(I + \frac{h}{2}A + \frac{h^2}{12}A^2). \qquad (1.4.31)$$

In [241], Man generalizes the above method to obtain a high-order iterative scheme, showing that the optimum order is two on the basis of computer time. The comparative table showing the number of multiplications necessary and storage needed for various numerical algorithms for solving (1.4.28) will be presented later.

An interesting class of iterative methods for solving (1.4.1) when A and B are stable, has been proposed by Hoskins, Meek and Walton [205] and it is extensively discussed in Hoskins, Meek and Walton [206,207,209], Hoskins, Pathan and Walton [210], Hoskins and Walton [211, 212] and Walton [274]. The essential idea behind this class of methods is as follows. Let, for k=0,1,2,...

$$A_{k+1} = \alpha_k A_k + \beta_k A_k^{-1}, \quad A_o=A \qquad (1.4.32)$$

$$B_{k+1} = \alpha_k B_k + \beta_k B_k^{-1}, \quad B_o=B \qquad (1.4.33)$$

and

$$C_{k+1} = \alpha_k C_k + \beta_k A_k^{-1} C_k B_k^{-1}, \quad C_o=C \qquad (1.4.34)$$

By induction, it follows that

$$A_k X + X B_k = C_k \text{ for } k=0,1,2,... \qquad (1.4.35)$$

The parameters α_k and β_k are chosen such that both A_k and B_k converge to $-I_m$ and $-I_n$ respectively. Consequently, the solution of (1.4.1) becomes

$$X = \lim_{k \to \infty} (-C_k/2). \qquad (1.4.36)$$

The classical choice of parameters is

$$\alpha_k = \beta_k = 1/2 \tag{1.4.37}$$

and this choice is associated with the well-known Newton's process. In order to accelerate the convergence it has been suggested [205,212] to take

$$\alpha_k = \frac{2a_k}{(1+\sqrt{a_k b_k})^2}, \quad \beta_k = a_k b_k \alpha_k \tag{1.4.38}$$

where

$$a_k = \min(\|A_k^{-1}\|^{-1}, \ \|B_k^{-1}\|^{-1}) \tag{1.4.39}$$

$$b_k = \max(\|A_k\|, \ \|B_k\|) \tag{1.4.40}$$

with any convenient norm. For the algorithm described by (1.4.32)-(1.4.37), the storage requirement is $2m^2+n^2+2mn$ for $m \geq n$ and m^2+2n^2+2mn otherwise; the number of multiplications required per iteration is approximately $\frac{4}{3}(m^3+n^3)+m^2n+mn^2$. It is claimed by Hoskins et al. [205] that this algorithm practically converges approximately in five iterations to achieve a solution with an accuracy of seven decimal places.

In [208], the above class of iterative procedure is considered for the Lyapunov matrix equation (1.4.28). In this case the step corresponding to (1.4.33) will not be there and in (1.4.34) and (1.4.35), B_k is to be replaced by A_k^T. In the same paper, the following choice of α_k and β_k has been proposed for accelerating the convergence:

$$\alpha_k = \frac{2a_k}{(a_k+\sqrt{a_k b_k})^2}, \quad \beta_k = a_k b_k \alpha_k \tag{1.4.41}$$

along with

$$a_{k+1} = 1-\varepsilon_k, \quad b_{k+1} = 1+\varepsilon_k \tag{1.4.42}$$

where

$$\varepsilon_k = \left\{\frac{a_k - \sqrt{a_k b_k}}{a_k + \sqrt{a_k b_k}}\right\}^2 \tag{1.4.43}$$

with

$$a_o = \|A^{-1}\|^{-1}, \quad b_o = \|A\|. \tag{1.4.44}$$

Pointing out that the method works for the last mentioned choice in general, only when the spectrum of A is real Barraud [163] has suggested the choice

$$\alpha_k = \frac{1}{2\sqrt{a_k b_k}}, \qquad \beta_k = \frac{1}{2}\sqrt{a_k b_k} \tag{1.4.45}$$

with

$$a_k = \|A_k^{-1}\|^{-1} \text{ and } b_k = \|A_k\| \tag{1.4.46}$$

in order to cover the complex spectrum case as well.

Earlier, Beavers and Denman [167] (also Denman and Beavers [54]) have described a method to solve the Lyapunov matrix equation (1.4.28) using the concept of matrix sign function which in fact is associated with an iterative scheme of the form

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-1}), \quad A_o = A. \tag{1.4.47}$$

In the course of solving the matrix differential equation

$$\dot{X} = AX + XB - C, \quad X(0) = Z \tag{1.4.48}$$

the solution of AX+XB=C is required [162,182,207,230]. In this context, Davison [182] extends the method of Davison and Man [183] based on the numerical quadrature to solve (1.4.1)

when A and B are stable and for the same case Hoskins, Meek and Walton [207] suggest the following values of α_k, β_k to implement the iterative scheme (1.4.32)-(1.4.36):

$$\alpha_k = \{tr(S_k)tr(S_k^{-2}) - p\ tr(S_k^{-1})\}/\gamma_k \qquad (1.4.49)$$

$$\beta_k = \{tr(S_k^2)tr(S_k^{-1}) - p\ tr(S_k)\}/\gamma_k \qquad (1.4.50)$$

$$\gamma_k = tr(S_k^2)tr(S_k^{-2}) - p^2 \qquad (1.4.51)$$

where

$$\left.\begin{aligned} &S_k=A_k \text{ and } p=m \text{ if } \rho(A_k) \geq \rho(B_k) \text{ and } \rho(A_k^{-1}) \geq \rho(B_k^{-1}) \\ &S_k=B_k \text{ and } p=n \text{ if } \rho(B_k) \geq \rho(A_k) \text{ and } \rho(B_k^{-1}) \geq \rho(A_k^{-1}) \end{aligned}\right\} (1.4.52)$$

and otherwise $\alpha_k = \beta_k = 1/2$.

Before noticing such a choice of α_k, β_k given in (1.4.49)-(1.4.51) we have developed independently such a choice with a slight modification and along with a generalization to solve the Lyapunov matrix equations (1.4.28) and (1.4.27). Moreover, by specializing to the Lyapunov matrix equation, it may be observed that the choice suggested by Hoskins et al. in (1.4.49)-(1.4.52) may not work in general when the spectrum of A is complex. To substantiate this statement we give theoretical counterexamples in Chapter 4 where we present our algorithms with detailed analysis.

Several papers are devoted to the comparative study of numerical methods for solving the Lyapunov matrix equation (1.4.28), for example, Bélanger and McGillivray [168], Hagander [197], Pace and Barnett [252] and Rothschild and Jameson [264]. In these papers, the algorithms adjudged

to be more efficient are those due to Bartels and Stewart [164], Jameson [215] and Smith [270]. Golub et al. [192] remark that the Hessenberg-Schur algorithm offers no advantage over the Bartels-Stewart method for the Lyapunov matrix equation.

Now we shall present the operation counts and storage needed in solving (1.4.28) by some of the standard methods that we have seen earlier. For the first four methods mentioned in the table, the matrix $A \in M_n(\mathbb{R})$ is assumed to be stable while for the last three methods this is not so. In the table, k denotes the number of iterations required for the method to converge numerically and the data have been collected from the literature.

TABLE 1.4.1
Comparison of numerical methods for solving the Lyapunov matrix equation

S.No.	Method	Approximate number of multiplications	Approximate storage locations
1.	Davison and Man [183]	$(2.5k+4)n^3$	$4n^2$
2.	Smith [270] (also refer [267])	$2.5(k+1)n^3$	$2.5n^2$
3.	Man [241]	$36.5n^3$	$4n^2$
4.	Hoskins, Meek and Walton [208]	$10n^3k/3$	$4n^2$
5.	Jameson [215] (also refer [264])	$O(n^4)$	not available
6.	Molinari [249]	$5n^3$	$4n^2$
7.	Bartels and Stewart [164]	$(2+4\sigma)n^3+3.5n^3$ where σ is the number of QR steps required.	$3n^2$

It seems that except probably in No.4 and 5, in all other methods C and X are assumed to be symmetric in calculating the number of operations.

Returning to some theoretical development on the Sylvester and Lyapunov equations, one can see expressions for bounds on solution of (1.4.1) in [231,270]. Bounds for the extreme eigenvalues of the solution matrix X of the Lyapunov matrix equations (1.4.27) and (1.4.28) for the case when A is stable and C is negative definite are discussed in [229,231,268].

The solvability of simultaneous Lyapunov matrix equations

$$AX + XA^* = XA + A^*X = I \tag{1.4.53}$$

where X is assumed to be Hermitian is a problem proposed by Taussky [82] and it has been studied by Davis [181], Gottlieb and Gunzburger [193] and Barker [150,151].

Finally, we shall mention a few words about the so-called discrete Lyapunov matrix equation

$$A^TXA - X = Q \tag{1.4.54}$$

associated with the linear discrete-time system

$$x_{k+1} = Ax_k. \tag{1.4.55}$$

It has been shown by Power [257] that the continuous and discrete Lyapunov matrix equations may be converted one to the other through the well-known Cayley transform referred in Section 1.2. The solution of (1.4.54) has important applications in the design of linear discrete systems [171].

Details of methods of solution of this equation may be found in Barnett [155], Barraud [161], Berger [171], Kitagawa [225], Power [258], Smith [269] and Young [277].

The equation

$$\Sigma \alpha_{ij} A^i X (A^T)^j = C \qquad (1.4.56)$$

which contains both the continuous and discrete Lyapunov matrix equations as special cases has been discussed by Barnett [156]. Techniques for solving the extended Lyapunov matrix equations

$$AX + XA^T - 2\sigma X = C \qquad (1.4.57)$$

and

$$\rho^{-2} AXA^T - X = C \qquad (1.4.58)$$

are presented by Heinen [203,204].

In the thesis, we also draw attention to certain projection and residual projection methods which can be effectively used for solving the Lyapunov and Sylvester equations even in singular cases.

2. NORMAL MATRICES AND ANGULARITY

2.1. Introduction

The purpose of this chapter is to introduce and study a notion of angularity as a generalization of the well-known concept of inertia of a matrix. The angularity characterizes the distribution of arguments of eigenvalues of a matrix and we give the formal definition of angularity in the section to follow.

In practical problems it is often insufficient to know merely that a linear system is asymptotically stable [2]. It is important to guarantee a certain degree of stability that can ensure a better performance of the transient process. This concept, known as relative stability [2,19] has mainly the following two specifications.

Suppose S_1 and S_2 are the two linear systems $\dot{x} = Ax$ and $\dot{x} = Bx$ respectively. If all the eigenvalues of A lie in the half plane $Re(z) < \alpha_1$ and those of B in the half plane $Re(z) < \alpha_2$, the system S_2 may be regarded as more stable than S_1 if $\alpha_2 < \alpha_1 < 0$. The second specification of relative stability is related to certain angular sectors of the complex plane. If all the eigenvalues of A lie in the sector $|\arg(-z)| < \beta_1$ and those of B in the sector $|\arg(-z)| < \beta_2$, the system S_2 may be regarded as having a better damping than S_1 if $0 \leq \beta_2 < \beta_1 < \pi/2$. In other words, the oscillations in the transient response will be less in S_2 than in S_1 (see Marden

relation with Hermitian matrices, almost all angularity theorems we prove in the thesis are concerned with normal matrices.

2.2. Angularity of a Matrix

We recall from Section 1.2 that the inertia $In(A)$ of $A \in M_n$, is the ordered triple $(\pi(A), \nu(A), \delta(A))$, the entries denoting the total number of eigenvalues of A with positive, negative, and zero real parts, respectively. Geometrically speaking, A has $\pi(A)$ eigenvalues in the open right half plane, $\nu(A)$ eigenvalues in the open left half plane and $\delta(A)$ eigenvalues on the imaginary axis, all counting multiplicities. The inertia $In(A)$ thus depends on the distribution of arguments of eigenvalues of A. This dependence is complete in the case of angularity $\theta[A]$ of A which we define in this section. Indeed, the aim of the present section is to explain and illustrate the concept of angularity of a matrix. In order to define $\theta[A]$, following Cain [36] we shall first define the ray space of the complex plane $\mathbb{C}$.

DEFINITION 2.2.1. The ray space of $\mathbb{C}$ is defined as the set $\Omega = \{\{0\}\} \cup \{e^{i\theta}\mathbb{R}_+ : 0 \leq \theta < 2\pi\}$ where $\mathbb{R}_+ = (0,\infty)$. A general element of Ω is called a ray and is denoted by ω, $\omega = \{0\}$ being called a null ray and $\omega = e^{i\theta}\mathbb{R}_+$ $(0 \leq \theta < 2\pi)$ being a proper ray.

It is understood that in $e^{i\theta}$, $i = \sqrt{-1}$ and θ will always be real. We note that the restriction $0 \leq \theta < 2\pi$ is of no

particular advantage in specifying the rays and hence in the sequel, the ray $\omega = e^{i(\theta+2k\pi)}\mathbb{R}_+$, where k is any integer will be identified with the ray $\omega = e^{i\theta}\mathbb{R}_+$.

By an extended ray we mean a straight line through the origin dividing the complex plane into two half planes. Clearly, an extended ray is the union of null ray and two proper rays with $\theta = \alpha$ and $\theta = \alpha+\pi$, α being arbitrary. We shall now introduce some notation.

$$\Omega_+ = \{e^{i\theta}\mathbb{R}_+ : |\theta| < \pi/2\}$$

$$\Omega_- = \{e^{i\theta}\mathbb{R}_+ : \pi/2 < \theta < 3\pi/2\}$$

$$\Omega_o = \Omega\setminus(\Omega_+ \cup \Omega_-)$$

$$\Omega_p = \Omega\setminus\{0\}.$$

Evidently the totality of the rays in Ω_+, Ω_-, Ω_o, and Ω_p refer respectively the open right half plane, open left half plane, imaginary axis and the origin deleted complex plane.

We can now give the definition of angularity of a matrix.

DEFINITION 2.2.2. The angularity $\theta[A]$ of $A \in M_n$ is a mapping from Ω to $\mathbb{N}$, the set of nonnegative integers for which $\theta[A]_\omega$ is the number of eigenvalues of A (counting multiplicities) lying on the ray ω.

This notion of angularity has been introduced independently in Cain [36] and Rathore and Chetty [75]. It may be noted that

$$\pi(A) = \sum_{\omega \in \Omega_+} \Theta[A]_\omega \qquad (2.2.1)$$

$$\nu(A) = \sum_{\omega \in \Omega_-} \Theta[A]_\omega \qquad (2.2.2)$$

$$\delta(A) = \sum_{\omega \in \Omega_0} \Theta[A]_\omega \,. \qquad (2.2.3)$$

Moreover A is stable iff $\Theta[A]_\omega = 0$ for all $\omega \in \Omega \setminus \underline{\Omega}$.

DEFINITION 2.2.3. Two matrices A and $B \in M_n$ are said to be equiangular iff $\Theta[A] = \Theta[B]$, i.e., $\Theta[A]_\omega = \Theta[B]_\omega$ for all $\omega \in \Omega$.

Obviously if A and B are equiangular then they have the same inertia, without the converse being true in general. For example, $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ have the same inertia. However, it is easily verified that they are not equiangular. In case if both A and B are Hermitian then they are equiangular iff they have the same inertia.

As we have pointed out earlier, we deal primarily with normal matrices in proving angularity theorems. Various types of matrices with which we are familiar are, in fact, normal. Among these are the diagonal matrices, Hermitian matrices, skew-Hermitian matrices, unitary matrices and circulants [4, p.242]. In this connection we shall find a necessary condition for two normal matrices to be equiangular. In Section 2.4, we show that this necessary condition is also sufficient.

THEOREM 2.2.1. Let A and B be two equiangular normal matrices. Then there exists a nonsingular matrix C such that $C^*AC = B$.

Proof. It is based on the spectral theorem. Let U and V be the unitary diagonalizers of the normal matrices A and B respectively. Since A and B are equiangular, without loss of generality (making use of a similarity transformation by a permutation matrix) we can assume that

$$U^*AU = \text{diag}(r_1 e^{i\alpha_1},\ldots,r_m e^{i\alpha_m},0,\ldots,0)$$

and

$$V^*BV = \text{diag}(R_1 e^{i\alpha_1},\ldots,R_m e^{i\alpha_m},0,\ldots,0)$$

with r_j, $R_j > 0$ and α_j real, $j=1,\ldots,m$. Let D be the diagonal matrix whose j-th entry is $(R_j/r_j)^{1/2}$, $j=1,\ldots,m$ and the remaining entries are unity. Then $D^*(U^*AU)D = V^*BV$. Hence with $C = UDV^{-1}$ we have $C^*AC = B$. This completes the proof.▲▲▲

We say that $\text{In}(A) \leq \text{In}(B)$ iff $\pi(A) \leq \pi(B)$ ane $\nu(A) \leq \nu(B)$. The same definition can be adopted even if A and B are of different size [47]. Now we shall define an analogous relation for the angularity of two matrices.

DEFINITION 2.2.4. If A and B are two square matrices, may be of different size, then we say that $\Theta[A] \leq \Theta[B]$ iff $\Theta[A]_\omega \leq \Theta[B]_\omega$ for all $\omega \in \Omega_p$.

From this definition it is clear that if $\Theta[A] \leq \Theta[B]$ then $\text{In}(A) \leq \text{In}(B)$. However, the converse is not true in general.

Finally, we list some simple observations regarding the notion of angularity of a matrix.

(i) $\theta[A] = \theta[A^T] = \theta[S^{-1}AS] = \theta[kA]$, k being a positive number.

(ii) $\theta[A^{-1}] = \theta[\bar{A}] = \theta[A^*]$, $\bar{A}$ denoting the conjugate of A.

(iii) Although $In(A) = In(A^{-1})$, it is not true in general that $\theta[A] = \theta[A^{-1}]$. However if A is real then $\theta[A] = \theta[A^{-1}]$.

(iv) We know that $In(A+A^{-1}) = In(A)$. However even for a real matrix such a relation is not true with the inertia replaced by angularity.

(v) $\theta[AB] = \theta[BA]$

(vi) If A is a quasi-diagonal matrix $diag(A_1,\ldots,A_m)$, Then $\theta[A]_\omega = \theta[A_1]_\omega + \ldots + \theta[A_m]_\omega$ for all $\omega \in \Omega$.

2.3. An Application of Angularity in Solving a Linear System

In the introductory section of this chapter we mentioned that a knowledge of the angularity of a matrix is useful in the study of relative stability and in reconstructing the eigenvalues in case their magnitudes are known. In the present section we indicate one more application of the notion of angularity in solving a linear system

$$Ax = b \qquad (2.3.1)$$

where A is an n x n non-Hermitian normal matrix.

Linear systems with this form of A occur in finite difference approximations to certain partial differential

equations [282], in the solution of transport equations by finite difference techniques [289], in solving banded Toeplitz matrices by circular decompositions [295] and possibly in many other applications.

To motivate considering an iterative method to solve (2.3.1), let us write it in the equivalent form

$$(A^*+A)x = (A^*-A)x + 2b. \tag{2.3.2}$$

This form suggests the iterative scheme

$$(A^*+A)x_{k+1} = (A^*-A)x_k + 2b \tag{2.3.3}$$

for solving (2.3.1). A necessary and sufficient condition for the convergence of this process may be expressed in terms of the angularity of A as in the following theorem.

THEOREM 2.3.1. Suppose A is normal such that $\delta(A)=0$ and $A \notin H_n$. Then for an arbitrary $b \in \mathbb{C}^n$ the iterative method

$$(A^*+A)x_{k+1} = (A^*-A)x_k + 2b$$

converges to the solution of $Ax=b$ for any choice of x_0 iff

$$\theta[A]_\omega = 0 \text{ for all } \omega \in \Omega\backslash(S_1 \cup S_2) \tag{2.3.4}$$

where

$$S_1 = \{e^{i\theta}\mathbb{R}_+ : 0 \leq |\theta| < \pi/4\} \tag{2.3.5}$$

and

$$S_2 = \{e^{i\theta}\mathbb{R}_+ : 0 \leq |\pi-\theta| < \pi/4\}. \tag{2.3.6}$$

Proof. Since A is normal and $\delta(A)=0$, A^*+A is nonsingular. Hence the given iterative scheme can be rewritten in the form

$$x_{k+1} = Bx_k + 2(A^*+A)^{-1}b \tag{2.3.7}$$

where $B = (A^*+A)^{-1}(A^*-A)$. It is well known that for an arbitrary $b \in \mathbb{C}^n$, (2.3.7) converges to the solution of $Ax=b$ for any initial x_o iff $\rho(B) < 1$. Furthermore we know that if the eigenvalues of A are $\lambda_1,\ldots,\lambda_n$ then the eigenvalues of B are $Im(-\lambda_i)/Re(\lambda_i)$, $i=1,\ldots,n$. From this it is readily seen that $\rho(B) < 1$ iff (2.3.4) holds. This completes the proof.▲▲▲

It may be noted that in the preceding theorem if A happens to be Hermitian then $B=0$ and hence there is no iterative character in the method proposed. Thus, if we know the angularity of A, assuming that A is non-Hermitian and normal, then we can decide whether to proceed or not iteratively as described above for solving $Ax=b$. This method is particularly advantageous when

$$\theta[A]_\omega = 0 \text{ for all } \omega \in \Omega \setminus S_1. \qquad (2.3.8)$$

In this case A^*+A is positive definite and hence can be factorized in the form LL^* where L is lower triangular. Therefore each step in (2.3.3) can be solved by the Cholesky method [21]. It may be noticed that the factorization LL^* has to be performed only once for the entire process, i.e. only for the first iteration. For the comparative study of the operation counts, let us further assume that the system $Ax=b$ is real. The proposed iterative-cum-Cholesky method requires n square roots, $(n^3+9n^2+2n)/6$ multiplications and $(n^3+6n^2-7n)/6$ additions in the first iteration [21]. These counts do not include the number of operations required in simplifying

the right side of (2.3.3). For this simplification we require n^2-n multiplications and n^2-n additions, noting that the diagonal elements of A^*-A are zero since A is assumed to be real. In fact, in the first iteration these operations do not come into the picture if we choose the initial vector as the zero vector.

It is not difficult to verify that each one of the successive iterations requires n^2-n multiplications and n^2-n additions in simplifying the right hand side vector; n^2+n multiplications and n^2-n additions in solving the two triangular systems involved in the Cholesky method. Thus the total number of operations for the proposed method comes to n square roots, $(n^3+9n^2+2n)/6 + 2(m-1)n^2$ multiplications and $(n^3+6n^2-7n)/6 + 2(m-1)(n^2-n)$ additions, m being the number of iterations required for the numerical convergence to take place. On the other hand, for solving Ax=b by Gaussian elimination, we require $(n^3+3n^2-n)/3$ multiplications and $(2n^3+3n^2-5n)/6$ additions. In practice, to compare the number of operations it is sufficient to consider only multiplicative operations. Thus we see that for large n, the proposed method requires about $n^3/6$ multiplications whereas Gaussian elimination requires about $n^3/3$ multiplications. This shows that the proposed method may still be better than Gaussian elimination for large n. Moreover no pivoting is needed in the Cholesky method [21].

In connection with the method suggested above for solving the linear system, it is desirable to find some necessary and sufficient condition for A to satisfy (2.3.8), i.e., for

all the eigenvalues of A to lie in the sector S_1 defined by (2.3.5). If A is normal then a necessary condition for A to satisfy (2.3.8) is that $A^*+A > 0$. However it is not sufficient. If A is nonnormal then the positive definiteness of A^*+A need not be even a necessary condition for A to satisfy (2.3.8). It is easily verified if we consider $A = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$.

If $A = (a_{ij})$ is real then a sufficient condition for A to have all the eigenvalues in S_1 is

$$a_{ii} > \sqrt{2} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i=1,\ldots,n. \tag{2.3.9}$$

However this is not necessary. This result is an immediate consequence of the fact that if (2.3.9) holds then all the n Gerŝgorin discs [16, p.146] of A lie in S_1.

In the next theorem, we give a condition which is necessary as well as sufficient for any general (need not be normal or real) n x n matrix to have the property (2.3.8).

THEOREM 2.3.2. Let $A \in M_n$. Then $\Theta[A]_\omega = 0$ for all $\omega \in \Omega\backslash S_1$ iff the inertia of 2n x 2n matrix $\begin{bmatrix} A & -A \\ A & A \end{bmatrix}$ is (2n,0,0) or, equivalently iff $\begin{bmatrix} -A & A \\ -A & -A \end{bmatrix}$ is stable.

Proof. Let $\lambda_1,\ldots,\lambda_n$ denote the characteristic values of A. We note that

$$\begin{bmatrix} A & -A \\ A & A \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \otimes A = C \text{ say,}$$

where $\otimes$ denotes the Kronecker product of matrices. Since the eigenvalues of $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ are $\sqrt{2}\exp(\pm i\pi/4)$ (the notation

$\exp(x)$ is often used in place of e^x, especially when x is a complicated expression), by a property mentioned earlier regarding the eigenvalues of the Kronecker product of two matrices, the eigenvalues of C are $\sqrt{2}\,\lambda_j \exp(\pm i\pi/4)$, $j=1,\ldots,n$. Consequently, we see that A satisfies (2.3.8) iff the eigenvalues of C lie in the open right half plane. This completes the proof of the theorem. ▲▲▲

In connection with Theorem 2.3.1, we shall now provide a simple necessary and sufficient condition for (2.3.4) to be true for any $A \in M_n$.

THEOREM 2.3.3. Let $A \in M_n$. Then $\theta[A]_\omega = 0$ for all $\omega \in \Omega\backslash(S_1 \cup S_2)$ iff $\mathrm{In}(A^2) = (n,0,0)$ or, equivalently iff A^2 is positive stable.

Proof. If $z \in \mathbb{C}$, $\mathrm{Re}(z^2) > 0$ iff either $0 \leq |\arg(z)| < \pi/4$ or $0 \leq |\pi - \arg(z)| < \pi/4$. The theorem is clear now. ▲▲▲

Finally, we note that the iterative scheme (2.3.3) may be rewritten in the equivalent form

$$x_{k+1} = x_k + ((A^*+A)/2)^{-1}(b-Ax_k). \qquad (2.3.10)$$

An implementation of (2.3.10) yields the residual $b-Ax_k$ as a bonus at each iteration. This could be useful for many purposes such as in deciding the accuracy of the solution or the termination criterion.

2.4. Angularity Theorems

In this section we are concerned with the angularity of

a normal matrix and that of a matrix congruent to A, i.e. a matrix of the form C^*AC where C is nonsingular. It is proved that if B and C are nonsingular then B^*AB and C^*AC are equiangular provided that the two matrices mentioned latter are normal. Some well-known inertia theorems (e.g. Sylvester's law of inertia) have been deduced as corollaries of this main result.

In order to prove our main result, we shall first prove a basic lemma. The method of proof closely resembles that used by Lancaster [13, p.89] to prove the classical version of Sylvester's law.

LEMMA 2.4.1. If B^*AB and C^*AC are diagonal matrices where B and C are nonsingular, then $\pi(C^*AC) = \pi(B^*AB)$.

Proof. Let $B^*AB = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and $C^*AC = \mathrm{diag}(\mu_1,\ldots,\mu_n)$ where without loss of generality we may assume that $\mathrm{Re}(\lambda_1) \geq \ldots \geq \mathrm{Re}(\lambda_n)$ and $\mathrm{Re}(\mu_1) \geq \ldots \geq \mathrm{Re}(\mu_n)$, $\mathrm{Re}(z)$ denoting the real part of z. Suppose that $\mathrm{Re}(\lambda_p)$, $\mathrm{Re}(\mu_q) > 0$ and $\mathrm{Re}(\lambda_{p+1})$, $\mathrm{Re}(\mu_{q+1}) \leq 0$. We have to prove that p=q.

Denoting the columns of B and C by $u_1,\ldots,u_n$ and $v_1,\ldots,v_n$ respectively, let us form the subspaces W_1 and W_2 of $\mathbb{C}^n$ spanned by $u_{p+1},\ldots,u_n$ and $v_1,\ldots,v_q$ respectively. If $q > p$, then $\dim(W_1) + \dim(W_2) = n-p+q > n$ where $\dim(W)$ stands for the dimension of the space W. Hence there exists a nonzero vector $x \in W_1 \cap W_2$. Let

$$x = \sum_{j=p+1}^{n} a_j u_j = \sum_{k=1}^{q} b_k v_k.$$

Then

$$\mathrm{Re}(x^*Ax) = \mathrm{Re}\Big(\sum_{j=p+1}^{n} |a_j|^2 \lambda_j\Big) \leq 0$$

and also

$$\mathrm{Re}(x^*Ax) = \mathrm{Re}\Big(\sum_{k=1}^{q} |b_k|^2 \mu_k\Big) > 0.$$

This contradiction leads to the conclusion $q \leq p$. By considering the subspaces spanned by $u_1,\ldots,u_p$ and $v_{q+1},\ldots,v_n$, a similar argument can be made to prove that $p \leq q$. Hence $p = q$. ▲▲▲

We next come to our principal result of this section.

THEOREM 2.4.1. If B^*AB and C^*AC are diagonal matrices where B and C are nonsingular, then $\theta[C^*AC] = \theta[B^*AB]$.

Proof. Let $A_\alpha = A \exp(i(\frac{\pi}{2} - \alpha))$, α being a real number. Applying Lemma 2.4.1 to the diagonal matrices $B^*e^{-i\varepsilon}A_\alpha B$ and $C^*e^{-i\varepsilon}A_\alpha C$ where $\varepsilon>0$, we find that $B^*A_\alpha B$ and $C^*A_\alpha C$ have the same number of eigenvalues in the open half plane $\{e^{i\theta}\mathbb{R}_+ : -\frac{\pi}{2} + \varepsilon < \theta < \frac{\pi}{2} + \varepsilon\}$. Moreover, these two matrices have the same number of eigenvalues in the open right half plane. Since we can choose ε sufficiently small such that both $B^*A_\alpha B$ and $C^*A_\alpha C$ have no eigenvalues in the regions $\{e^{i\theta}\mathbb{R}_+ : -\frac{\pi}{2} < \theta \leq -\frac{\pi}{2} + \varepsilon\}$ and $\{e^{i\theta}\mathbb{R}_+ : \frac{\pi}{2} < \theta < \frac{\pi}{2} + \varepsilon\}$ we see that the two diagonal matrices $B^*A_\alpha B$ and $C^*A_\alpha C$ will have the same number of eigenvalues on the open upper half of the imaginary axis. But this is equivalent to say that B^*AB and C^*AC have the same number of eigenvalues on the ray $\{e^{i\alpha}\mathbb{R}_+\}$. Since α is arbitrary the theorem follows. ▲▲▲

Next, suppose that B^*AB and C^*AC are normal matrices where B and C are nonsingular. If U and V are the respective unitary diagonalizers of these normal matrices then by the above theorem $\theta[V^*C^*ACV] = \theta[U^*B^*ABU]$ which implies that C^*AC and B^*AB are equiangular. Thus we arrive at

COROLLARY 2.4.1. If B^*AB and C^*AC are normal matrices where B and C are nonsingular, then $\theta[C^*AC] = \theta[B^*AB]$.

Also, we have the following two immediate corollaries.

COROLLARY 2.4.2. If A and C^*AC are diagonal where C is nonsingular, then $\theta[C^*AC] = \theta[A]$.

COROLLARY 2.4.3. If A and C^*AC are normal where C is nonsingular, then $\theta[C^*AC] = \theta[A]$.

Another interesting corollary in this sequence is

COROLLARY 2.4.4. If A and C are circulants and C is nonsingular then $\theta[C^*AC] = \theta[A]$.

The last corollary is an immediate consequence of the fact that the product of circulants is again a circulant and the circulants are always normal [13, p.267]. We will study circulants in detail in the next chapter.

REMARK 2.4.1. One can readily prove that Theorem 2.4.1 and its first three corollaries are all equivalent to one another.

We have already shown that Theorem 2.4.1 ==> Corollary 2.4.1. Obviously Corollary 2.4.1 ==> Corollary 2.4.3. Since diagonal matrices are always normal, we have Corollary 2.4.3 ==> Corollary 2.4.2. Finally, if B^*AB and C^*AC are diagonal

then by viewing C^*AC as $(B^{-1}C)^*B^*AB(B^{-1}C)$ it follows that Corollary 2.4.2 ==> Theorem 2.4.1. This completes the cycle.

REMARK 2.4.2. Using the polar decomposition of C, Corollary 2.4.3 has been proved independently by Cain [36, Theorem 6.5].

Combining Theorem 2.2.1 and Corollary 2.4.3 we have the following theorem which gives the necessary and sufficient condition for two normal matrices to be equiangular.

THEOREM 2.4.2. Two given normal matrices A and B are equiangular, iff there exists a nonsingular matrix C such that $C^*AC = B$.

We have seen, in Corollary 2.4.3, that if A and C^*AC are normal, C being nonsingular, then they are equiangular. If the normality restriction on C^*AC is removed in this, then $\theta[C^*AC]$ need not equal $\theta[A]$ as may be easily verified by taking $A = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}$.

In this connection, we ask ourselves the following question. Let A be normal and C be nonsingular such that $\theta[C^*AC] = \theta[A]$. Does it imply that C^*AC is normal? Alternatively is there any nonsingular C such that C^*AC is nonnormal, A normal and $\theta[C^*AC] = \theta[A]$? Let us put this question in another simple form without involving angularity explicitly.

Suppose A is normal, C is nonsingular and C^*AC and A are equiangular. Let U be the unitary diagonalizer of A so that $U^*AU = D$. By Schur's theorem [4, p.202], there exists a unitary matrix V such that $V^*(C^*AC)V = T$ where T is an upper triangular matrix. It is well known that C^*AC is normal iff T is diagonal. If T is expressed as $\tilde{D} + \tilde{T}$ where $\tilde{D}$ is diagonal and $\tilde{T}$ is strictly upper triangular, then from the hypothesis $\tilde{D}$ and D are equiangular. Hence by virtue of Theorem 2.4.2, there exists a nonsingular matrix W such that $W^*\tilde{D}W = D$. Now we have $V^*C^*UW^*\tilde{D}WU^*CV = \tilde{D} + \tilde{T}$. Writing WU^*CV as M, it reduces $M^*\tilde{D}M = \tilde{D} + \tilde{T}$. In view of this argument, the problem proposed in the last paragraph may be stated as follows:

Let D be a diagonal matrix and C be nonsingular such that

$$C^*DC = D + T, \qquad (2.4.1)$$

T being strictly upper triangular. Does it imply that $T = 0$? If either C or D is real then it can be easily shown that $T = 0$. If C is real then $C^* = C'$. Hence by taking transpose on both sides of (2.4.1), we have $T = T'$ implying $T = 0$. If D is real then by taking conjugate transpose on both sides of (2.4.1), it follows $T = T^*$ and hence $T = 0$. Therefore if $T \neq 0$, then necessarily both C and D should be complex. It is not known to us whether in general (2.4.1) implies $T = 0$. However in the following theorem we are able to answer affirmatively our question for the case n=2. For $n \geq 3$, the problem remains open.

THEOREM 2.4.3. Suppose D is a 2x2 diagonal matrix and C is a 2x2 nonsingular matrix such that $C^*DC = D + T$, T being strictly upper triangular. Then T = 0.

Proof. Let

$$C = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \text{ and } T = \begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix}.$$

Hence from $C^*DC = D + T$, we have

$$k\begin{bmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{bmatrix} \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} = \begin{bmatrix} d_1 & t \\ 0 & d_2 \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{2.4.2}$$

where $k = \det(C) = ad-bc$. From (2.4.2) we get

$$k\bar{a}d_1 = dd_1 - ct \tag{2.4.3}$$

$$k\bar{c}d_2 = at - bd_1 \tag{2.4.4}$$

$$k\bar{b}d_1 = -cd_2 \tag{2.4.5}$$

$$k\bar{d}d_2 = ad_2. \tag{2.4.6}$$

From (2.4.3) and (2.4.4) or directly from $C^*DC = D + T$, we also get

$$t = \bar{a}bd_1 + \bar{c}dd_2. \tag{2.4.7}$$

Our aim is to prove that t=0 in any case. Now let us consider the case $d_1=0$. Then from (2.4.3) it follows that c=0 or t=0. If c=0 then from (2.4.7), t=0. If $d_2=0$, by (2.4.5), $\bar{b}d_1=0$, which by (2.4.7) implies that t=0. So we assume that $d_1d_2 \neq 0$. Consequently, we see that $|k|=1$. Now from (2.4.6), $a=k\bar{d}$. Substituting this in (2.4.3) gives that c=0 or t=0. If c=0 then from (2.4.5), b=0 and hence from (2.4.7), t=0. Thus in any case t=0. This completes the proof. ▲▲▲

In view of the above theorem and Theorem 2.4.1, we have thus established for 2x2 matrices the following result.

THEOREM 2.4.4. Let A be normal and C be nonsingular. Then C^*AC and A are equiangular iff C^*AC is normal.

For matrices in M_n, $n \geq 3$ one part is always true and it will be interesting to investigate the other part also.

Since the normality of C^*AC is of fundamental importance in our angularity theorems, we proceed now to derive some results involving the normality of C^*AC. As a first result, we shall now characterize the class of all nonsingular matrices C for which C^*AC is normal for every normal matrix A. This result will also find an application in proving one of our main theorems regarding the angularity-preserving linear transformations in the forthcoming chapter.

THEOREM 2.4.5. For a nonsingular C, C^*AC is normal for every normal matrix A, iff C is a nonzero scalar multiple of a unitary matrix.

Proof. The "if" part is obvious. Conversely, if C^*AC is normal for every normal A, then $ACC^*A^* = A^*CC^*A$ for every normal A. We may choose A to be a diagonal matrix with any one of the diagonal entries as 1 and the remaining (n-1) diagonal entries as i $(= \sqrt{-1})$, to show that CC^* is diagonal. Further, if we choose A as the matrix $E_{jj}+E_{kj}+E_{kk}-E_{jk}$, $j \neq k$ where E_{rs} denotes the matrix whose (r,s) element is unity and all other elements are zero, it is easily seen that the j-th and k-th diagonal entries of CC^* are equal. This holds for every pair of j and k. Hence, noting that $CC^* > 0$, we have

$CC^* = \alpha I$, $\alpha > 0$. The theorem now follows immediately. ▲▲▲

In the next theorem we give a sufficient condition for C^*AC to be normal when A is normal.

THEOREM 2.4.6. Let A be normal. Then C^*AC is normal if A commutes with CC^*.

Proof. Following Bellman [4, p.27], let us introduce the Jacobi bracket symbol $[A,B]$ to denote $AB-BA$, the commutator of A and B. By definition, A is normal iff $[A,A^*] = 0$. Now if we write $P = CC^*$ we have $AP=PA$ and also $PA^*=A^*P$ since $P^*=P$. By simple calculations we have $[C^*AC, C^*A^*C] = C^*P[A,A^*]C = 0$ and the result follows. ▲▲▲

We remark that the above result holds even if C is singular or in general if C is an $n \times m$ matrix with, of course, A in M_n. We study the angularity theorems with C as singular or rectangular in the following section.

We also remark that the commutativity of A and CC^*, however, is not necessary for C^*AC to be normal if A is given to be normal. A simple example to see this is $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

As an immediate consequence of Theorem 2.4.6 we have

THEOREM 2.4.7. Let A be normal and P be positive definite such that $PA = AP$. Then for any C satisfying $CC^* = P$, $\theta[C^*AC] = \theta[A]$.

We next turn to the angularity analogue of Sylvester's law

of inertia. We have already seen the importance and various forms of this classical law in the review chapter. In literature, this law is usually associated with Hermitian matrices only. In the course of studying its angularity analogue, we propose, now, to extend this association to a generalized class of Hermitian matrices. This consists of the so-called co-Hermitian matrices introduced by Ballantine [280,281] and defined as follows.

DEFINITION 2.4.1. A matrix $K \in M_n$ is said to be co-Hermitian iff $e^{-i\theta}K$ is Hermitian for some real θ.

Equivalently $K \in M_n$ is co-Hermitian iff $K = zH$ for some nonzero complex number z and $H \in H_n$. We shall denote the class of all $n \times n$ co-Hermitian matrices by K_n. Evidently, K_n contains H_n as well as the class of all $n \times n$ skew-Hermitian matrices. Another simple, but interesting observation is that co-Hermitian matrices are always normal.

Before coming to the proposed study of angularity analogue of Sylvester's law, we shall give some characterizations of co-Hermitian matrices.

LEMMA 2.4.2. A normal matrix $K \in M_n$ is co-Hermitian iff all the characteristic roots of K lie on an extended ray of the complex plane.

Proof. This is an immediate consequence of the well-known result about Hermitian matrices [16, p.64], that a normal matrix A is Hermitian iff the characteristic roots of A are all real. ▲▲▲

LEMMA 2.4.3. K is co-Hermitian iff $K^* = \alpha K$ for some complex number α such that $|\alpha|=1$.

Proof. This result is given in Ballantine [281]. If $K^* = \alpha K$, then by expressing α as $\exp(i\beta)$, we find that $\exp(i\beta/2)K$ is Hermitian. Hence K is co-Hermitian.

On the other hand, if $K = zH$ for $z\ (\neq 0) \in \mathbb{C}$ and $H \in H_n$ we have $K^* = (\bar{z}/z)K$. Hence $K^* = \alpha K$ with $|\alpha|=1$. ▲▲▲

In Theorem 2.4.5 we characterized the class of all nonsingular matrices C such that C^*AC is normal for every normal matrix A. Now in the following lemma, we shall characterize the class of all matrices A such that C^*AC is normal for every nonsingular matrix C. The answer turns out to be the class of all co-Hermitian matrices.

LEMMA 2.4.4. Let $A \in M_n$. Then C^*AC is normal for every nonsingular C iff A is co-Hermitian.

Proof. Suppose $A \in K_n$. Then $A = zH$ for $z\ (\neq 0) \in \mathbb{C}$ and $H \in H_n$. By straightforward calculations it follows that $[C^*AC, C^*A^*C] = 0$, completing the sufficiency part of the lemma.

Next, suppose that C^*AC is normal for every nonsingular C. Therefore, we immediately infer that A is normal. Let $U^*AU = D$, where U is unitary and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. For the choice $C = U(I+J)$ where I is the identity matrix and J is the matrix of unities, i.e., the matrix having all its elements as 1, the normality of C^*AC leads to the relation

$$D(I+J)^2D^* = D^*(I+J)^2D$$

which in view of the fact $J^2 = nJ$ further simplifies to $DJD^* = D^*JD$. From this we get

$$\lambda_i \bar{\lambda}_j = \bar{\lambda}_i \lambda_j \text{ for } i,j=1,\ldots,n.$$

and consequently $\bar{\lambda}_i = \alpha\lambda_i$ for $i=1,\ldots,n$, α being a constant such that $|\alpha|=1$. Hence we have $D^* = \alpha D$ and therefore, $A^* = \alpha A$ which completes the proof of the lemma by virtue of the preceding lemma. ▲▲▲

We shall now study the angularity analogue of Sylvester's law. The first assertion of the following result can be considered as the angularity analogue of Sylvester's inertia theorem.

THEOREM 2.4.8. The following three statements are true and are equivalent to each other:

(i) If C is nonsingular and $K \in K_n$, then $\theta[C^*KC] = \theta[K]$.

(ii) If $P > 0$ and $K \in K_n$, then $\theta[PK] = \theta[K]$.

(iii) If $K \in K_n$ and $RK > 0$, then $\theta[R] = \theta[K^*]$.

Proof. (i) is a consequence of Corollary 2.4.3. To prove (i) ==> (ii), as $P > 0$, it has a positive definite Hermitian square root $P^{1/2}$. Thus $\theta[PK] = \theta[P^{1/2}P^{1/2}K] = \theta[P^{1/2}KP^{1/2}] = \theta[K]$. In view of the positive definiteness of CC^*, (ii) ==> (i).

As $RK > 0$, K is nonsingular and hence $\theta[R] = \theta[RKK^{-1}] = \theta[K^{-1}] = \theta[K^*]$ showing that (ii) ==> (iii). The converse follows in a straightforward manner if one carries out the steps as in the proof of Corollary 3 of Ostrowski and Schneider

[71]. However, for the sake of completeness we shall give the proof. If K is nonsingular, then $P = PKK^{-1} > 0$ and hence by (iii) $\theta[PK] = \theta[K^{-*}] = \theta[K]$ implying (ii) in this case. Here and in the sequel the notation K^{-*} is used to denote $(K^*)^{-1}$ which is also $(K^{-1})^*$.

If K is singular, we choose a unitary U so that $U^*KU = D = \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix}$ in partition form where D_1 is a nonsingular diagonal matrix, noting that K is normal. Partition $Q = U^*PU$ as $\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ so that $QD = U^*(PK)U = \begin{bmatrix} Q_{11}D_1 & 0 \\ Q_{21}D_1 & 0 \end{bmatrix}$. We have to prove that $\theta[PK] = \theta[K]$ i.e., $\theta[QD] = \theta[D]$. To establish this it is enough to show that $\theta[Q_{11}D_1] = \theta[D_1]$. Since $D_1^{-1} \in K_n$, and $(Q_{11}D_1)D_1^{-1} = Q_{11} > 0$, by (iii) $\theta[Q_{11}D_1] = \theta[D_1^{-*}] = \theta[D_1]$. The proof of the theorem is therefore complete. ▲▲▲

Recently, it has been reported by Cain [36] that D.G.Hook, in an unpublished thesis written under C.S.Ballantine, has given a nice characterization of matrices of Sylvester type, i.e. matrices A satisfying $In(C^*AC) = In(A)$ for every invertible C. The result is as follows:

THEOREM 2.4.9 [36]. $In(C^*AC) = In(A)$ for every invertible C iff the numerical range (or the field of values) of A,

$$W(A) = \{x^*Ax: x^*x = 1\}$$

lies either (i) in a straight line through 0 or (ii) in the open right half plane together with 0 or (iii) in the open left half plane together with 0.

Motivated by this result, we characterize below, the class of all matrices A satisfying $\Theta[C^*AC] = \Theta[A]$ for every invertible C. Once again, as in Lemma 2.4.4, the answer turns out to be the class K_n.

THEOREM 2.4.10. Let $A \in M_n$. Then $\Theta[C^*AC] = \Theta[A]$ for every nonsingular C iff $A \in K_n$.

Proof. Theorem 2.4.8(i) furnishes the proof of "if" part. Conversely, let $\Theta[C^*AC] = \Theta[A]$ for every nonsingular C. From this, it is clear that $\mathrm{In}(C^*e^{i\alpha}AC) = \mathrm{In}(e^{i\alpha}A)$ for every nonsingular C where α is an arbitrary real number. By the above theorem, we see that for each real α, $W(e^{i\alpha}A)$, i.e., $e^{i\alpha}W(A)$ lies in any one of the three sets mentioned in the statement of that theorem.

Our claim is that for every α, the first possibility holds. If it is not so, then for $\alpha = \alpha_o$ say, there will exist z_1 $(= r_1\exp(i\phi_1))$ and z_2 $(= r_2\exp(i\phi_2))$ with r_1, r_2, $\mathrm{Re}(z_1)$, $\mathrm{Re}(z_2) > 0$ and $0 < (\phi_1-\phi_2) < \pi$ such that $z_1, z_2 \in \exp(i\alpha_o)W(A)$. Consequently, for

$$\beta = \alpha_o + \{\pi-(\phi_1+\phi_2)\}/2,$$

$\tilde{z}_1, \tilde{z}_2 \in e^{i\beta}W(A)$ where

$$\tilde{z}_1 = r_1\exp\{(\pi+\phi_1-\phi_2)i/2\}$$

and

$$\tilde{z}_2 = r_2\exp\{(\pi+\phi_2-\phi_1)i/2\}.$$

It can be easily verified that

$$\mathrm{Re}(\tilde{z}_1)\ \mathrm{Re}(\tilde{z}_2) < 0.$$

Moreover $\arg(\tilde{z}_1)-\arg(\tilde{z}_2)$ is not a multiple of π. Hence for $\alpha = \beta$, $e^{i\alpha}W(A)$ does not satisfy any one of the three conditions mentioned earlier. It is a contradiction. Hence for every α, $e^{i\alpha}W(A)$ lies in a straight line through 0. In particular, $W(A)$ lies in an extended ray which in turn implies that for some real ϕ, $x^*(e^{-i\phi}A)x$ is real for all $x \in \mathbb{C}^n$ such that $x^*x = 1$. Consequently, we have that $e^{-i\phi}A = e^{i\phi}A^*$. Hence by Lemma 2.4.3, it follows that $A \in K_n$. ▲▲▲

The argument just used is a consequence of the fact that if x^*Ax is real for all $x \in \mathbb{C}^n$, then A is Hermitian. We may also note that A is Hermitian iff $W(A)$ is an interval on the real line [16, p.169]. Based on this we can conclude that A is co-Hermitian iff $W(A)$ is an interval on an extended ray.

We may summarize the conclusions reached so far about co-Hermitian matrices in the form of a theorem.

THEOREM 2.4.11. Let $A \in M_n$. Then the following statements are equivalent.

(i) $A \in K_n$.

(ii) A is normal and has all its eigenvalues on an extended ray.

(iii) $A^* = \alpha A$ for some complex number α such that $|\alpha|=1$.

(iv) C^*AC is normal for every nonsingular matrix C.

(v) $\Theta[C^*AC] = \Theta[A]$ for every nonsingular matrix C.

(vi) $W(A)$ is an interval on an extended ray.

Since $\Theta[A] = \Theta[B]$ implies that $In(A) = In(B)$, the angularity results proved in this section, namely Theorem 2.4.1, its four corollaries, Theorems 2.4.7 and 2.4.8 may be restated as inertia theorems by replacing "Θ" by "In". For instance, one such a result is

THEOREM 2.4.12. If A and C^*AC are normal and C is nonsingular then $In(C^*AC) = In(A)$.

This may be regarded as Sylvester's inertia theorem for normal matrices [36]. We sketch an alternative proof of this theorem below, by using the classical inertia theorem due to Sylvester. It is well known that if A is normal then $In(A+A^*) = In(A)$. Applying this to C^*AC, $In(C^*AC) = In(C^*AC+C^*A^*C) = In(C^*(A+A^*)C)$. Since C is nonsingular and $A+A^*$ is Hermitian, we have $In(C^*(A+A^*)C) = In(A+A^*) = In(A)$ completing the proof. ▲▲▲

A quantitative sharpening of this result will be given later in Section 2.6 where we discuss some more generalizations of Sylvester's law based on the work of Ostrowski [69,70] and Thompson [85,86].

The following theorem is the inertia analogue of Theorem 2.4.2.

THEOREM 2.4.13. Two given normal matrices A and B have the same inertia, iff there exists a nonsingular matrix C such that $C^*Re(A)C = Re(B)$.

Proof. Again, we employ the result that $In(Re(A)) = In(A)$ when A is normal. By means of this result and Sylvester's theorem the sufficiency part follows immediately.

To prove the other part, we note that $In(A) = In(B)$ implies $In(Re(A)) = In(Re(B))$ and hence also $\theta[Re(A)] = \theta[Re(B)]$, since for Hermitian matrices equi-inertia property implies equiangularity. Therefore by Theorem 2.2.1 there exists a nonsingular C as required. ▲▲▲

We have seen that the term "inertia" can be simply replaced by "angularity" in Sylvester's theorem. However this is not possible with all inertia theorems. For example, if we let $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, then $Re(AH) > 0$, but $\theta[A] \neq \theta[H]$. This shows that "inertia" cannot be replaced by "angularity" in the main inertia theorem (refer Theorem 1.2.3).

2.5. Angularity Theorems in Singular Case

Referring to the results established in the preceding section, we now pass on to a discussion of the case where C is permitted to be singular and accordingly the concerned positive definite matrices are assumed to be positive semidefinite. Also, we extend our angularity results for the case when C is a rectangular matrix of order $n \times m$.

The first result in this sequence is the one corresponding to the basic lemma of the last section.

LEMMA 2.5.1. If B^*AB and C^*AC are diagonal matrices, and B is nonsingular, then $\pi(C^*AC) \leq \pi(B^*AB)$.

Proof. With reference to the notation and convention used in the proof of Lemma 2.4.1, first we shall prove that the first q columns $v_1,\ldots,v_q$ of C are linearly independent even if C is singular. To prove this, let us assume that

$$\sum_{k=1}^{q} c_k v_k = 0 \tag{2.5.1}$$

where $c_1,\ldots,c_q$ are scalars. We claim that $c_k=0$ for $k=1,\ldots,q$. From (2.5.1), we have

$$(\sum_{k=1}^{q} c_k v_k)^* A (\sum_{k=1}^{q} c_k v_k) = 0. \tag{2.5.2}$$

It is not difficult to see that $v_i^* A v_j$ is the (i,j) element of C^*AC. Since C^*AC is diagonal, (2.5.2) reduces to

$$\sum_{k=1}^{q} |c_k|^2 \mu_k = 0 \tag{2.5.3}$$

and hence

$$\sum_{k=1}^{q} |c_k|^2 \mathrm{Re}(\mu_k) = 0. \tag{2.5.4}$$

Since $\mathrm{Re}(\mu_k) > 0$ for $k=1,\ldots,q$, it follows that $c_k=0$ for $k=1,\ldots,q$. We may now observe that the first half of the analysis of the proof of Lemma 2.4.1 constitutes the proof of this lemma. ▲▲▲

By an arbitrary open half plane we mean a set of the form $\{e^{i\theta}\mathbb{R}_+ : \alpha < \theta < \alpha+\pi\}$, α being an arbitrary real number.

Assuming that B^*AB and C^*AC are diagonal and B is nonsingular, let us apply the above lemma to the diagonal matrices $B^*e^{i\alpha}AB$ and $C^*e^{i\alpha}AC$ where α is real. As a result,

we infer that in any arbitrary open half plane, the number of eigenvalues of C^*AC does not exceed the number of eigenvalues of B^*AB in that half plane. This may be expressed in the language of angularity as follows:

THEOREM 2.5.1. If B^*AB and C^*AC are diagonal and B is nonsingular, then

$$\sum_{\omega \in S} \theta[C^*AC]_\omega \leq \sum_{\omega \in S} \theta[B^*AB]_\omega \ , \tag{2.5.5}$$

S being any open half plane.

Corollaries 2.4.1-2.4.4 have similar analogues.

An interesting feature of our principal result, namely Theorem 2.4.1 of the preceding section lies in the following fact. In any arbitrary open half plane if two matrices have equal number of eigenvalues then on any arbitrary proper ray also they have equal number of eigenvalues. Based on this, it is now tempting to conclude under the assumptions of Theorem 2.5.1 that

$$\theta[C^*AC]_\omega \leq \theta[B^*AB]_\omega \text{ for all } \omega \in \Omega_p,$$

that is,

$$\theta[C^*AC] \leq \theta[B^*AB] \tag{2.5.6}$$

in view of Definition 2.2.4. Unfortunately it is not true. To substantiate this, a counterexample is given below. Let

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

so that

$$C^*AC = \text{diag}(1,0) \text{ and } B^*AB = \text{diag}(2+2i, 2-2i).$$

Evidently, for $\omega = \mathbb{R}_+$, $\Theta[C^*AC]_\omega = 1$ and $\Theta[B^*AB]_\omega = 0$ showing that (2.5.6) does not hold.

Therefore, when C is permitted to be singular "=" cannot be replaced by "$\leq$" in Theorem 2.4.1 and its four corollaries. However such a replacement is valid in the case of angularity analogue of Sylvester's theorem when C is permitted to be singular. To illustrate, we state the counterparts of Theorem 2.4.8(i) and (ii) in the singular and semidefinite case as:

THEOREM 2.5.2. The following two statements are true and are equivalent to each other.

(i) If $C \in M_n$ and $K \in K_n$, then $\Theta[C^*KC] \leq \Theta[K]$.

(ii) If $P \geq 0$ and $K \in K_n$, then $\Theta[PK] \leq \Theta[K]$.

Proof. Since K and C^*KC are normal, by the earlier theorem we find that in any arbitrary open half plane, the number of eigenvalues of C^*KC does not exceed the number of eigenvalues of K in that half plane. Moreover both K and C^*KC have their eigenvalues on one and the same extended ray. Hence, it is clear that $\Theta[C^*KC]_\omega \leq \Theta[K]_\omega$ for all $\omega \in \Omega_p$. This completes the proof of assertion (i)

(i) <==> (ii) follows easily as in Theorem 2.4.8. ▲▲▲

Corresponding to the third statement in Theorem 2.4.8, we have the following analogue which requires the assumption that K be nonsingular and it follows from Theorem 2.5.2(ii).

THEOREM 2.5.3. If $K \in K_n$ is nonsingular and $PK \geq 0$ then $\theta[P] \leq \theta[K^*]$.

In view of (2.2.1) and (2.2.2) the results of this section easily lead to the corresponding inertia theorems given below.

THEOREM 2.5.4. The following two statements are true and are equivalent.

(i) If $C \in M_n$ and $K \in K_n$, then $In(C^*KC) \leq In(K)$.

(ii) If $P \geq 0$ and $K \in K_n$, then $In(PK) \leq In(K)$.

In this theorem, in particular, if K is considered as Hermitian then (i) and (ii) are respectively Theorem 2 in Ostrowski [69] and a special case of Corollary 4 in Ostrowski and Schneider [71].

Since $In(K^*) = In(K)$, the inertia analogue to Theorem 2.5.3 becomes

THEOREM 2.5.5. If $K \in K_n$ is nonsingular and $PK \geq 0$, then $In(P) \leq In(K)$.

This may be compared with a particular case of a part of Lemma 2 in Carlson and Schneider [47].

So far we have confined ourselves to the simple situation in which C is a square matrix. Now, we will drop this assumption and assume instead that C is an $n \times m$ matrix with $m >$, $=$, or, $< n$. Of course, the other matrices concerned are in M_n. It can be easily seen that all the results involving C proved in this section remain valid.

Once we prove that Lemma 2.5.1 holds in rectangular case, then other results follow automatically since the same arguments can be applied as in the square case. We will therefore have to prove

THEOREM 2.5.6. Suppose $C \in M_{n,m}$, A and $B \in M_n$ and B is nonsingular. If B^*AB and C^*AC are diagonal matrices then $\pi(C^*AC) \leq \pi(B^*AB)$.

Proof. The proof of Lemma 2.5.1 remains valid without change for this case also. However, by assuming that the result is valid for square case, we complete the proof of this theorem by making use of a simple idea employed by Ostrowski [70] while extending the quantitative formulation of Sylvester's law of inertia to the rectangular case.

We have to consider three cases. The case m=n is simply Lemma 2.5.1. Now we consider the case $m < n$. Let us form $\tilde{C} \in M_n$ by augmenting (n-m) columns, consisting of zeros to C so that $\tilde{C} = [C\ 0]$. Since $\tilde{C}^*A\tilde{C} = \begin{bmatrix} C^*AC & 0 \\ 0 & 0 \end{bmatrix}$, it follows that $\pi(C^*AC) = \pi(\tilde{C}^*A\tilde{C}) \leq \pi(B^*AB)$.

Assuming $m > n$, now let us form the m x m matrices $\tilde{A} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$, $\tilde{B} = \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix}$, $\tilde{C} = \begin{bmatrix} C \\ 0 \end{bmatrix}$ where the block matrices are of appropriate dimensions. Since $\tilde{C}^*\tilde{A}\tilde{C} = C^*AC$, $\tilde{B}^*\tilde{A}\tilde{B} = \begin{bmatrix} B^*AB & 0 \\ 0 & 0 \end{bmatrix}$ and B is nonsingular, it follows that $\pi(C^*AC) = \pi(\tilde{C}^*\tilde{A}\tilde{C}) \leq \pi(\tilde{B}^*\tilde{A}\tilde{B}) = \pi(B^*AB)$. ▲▲▲

The inequality just established concerning $\pi(C^*AC)$ and $\pi(B^*AB)$ can further be generalized involving the rank of C.

For this we require the following lemma.

LEMMA 2.5.2. Let K be a principal submatrix of order r of $H \in H_n$. Then

$$\max(0, \pi(H)-n+r) \leq \pi(K) \leq \min(r, \pi(H)). \qquad (2.5.7)$$

Proof. This is an immediate consequence of the well-known interlacing Cauchy's inequalities [16, p.119] for the eigenvalues of principal submatrices of Hermitian matrices. If $\alpha_1 \geq \ldots \geq \alpha_n$ are the eigenvalues of H and $\beta_1 \geq \ldots \geq \beta_{n-1}$ are the eigenvalues of a principal submatrix G or order n-1 of H, then

$$\alpha_i \geq \beta_i \geq \alpha_{i+1}, \quad i=1,\ldots,n-1 \qquad (2.5.8)$$

so that

$$\pi(H)-1 \leq \pi(G) \leq \pi(H). \qquad (2.5.9)$$

By a repeated application of (2.5.9) we have,

$$\pi(H)-n+r \leq \pi(K) \leq \pi(H). \qquad (2.5.10)$$

Furthermore,

$$0 \leq \pi(K) \leq r. \qquad (2.5.11)$$

From (2.5.10) and (2.5.11) the required result follows. ▲▲▲

THEOREM 2.5.7. Let C be a matrix of order nxm and of rank r. If B^*AB and C^*AC are diagonal and B is invertible then

$$\max(0, \pi(B^*AB)-n+r) \leq \pi(C^*AC) \leq \min(r, \pi(B^*AB)) \qquad (2.5.12)$$

Proof. Since C is of rank r, there exist nonsingular matrices X and Y (of order n, m respectively) such that

$$C = X \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Y$$

where the block null matrices are of appropriate order [18, p.177]. Let $H = A+A^*$ and $\tilde{H} = X^*HX$. If $\tilde{H}$ is written in the partitioned form $(\tilde{H}_{ij})_{i,j=1,2}$ where $\tilde{H}_{11} \in M_r$, $\tilde{H}_{22} \in M_{n-r}$, then by usual simplifications

$$\pi(C^*HC) = \pi(\tilde{H}_{11}). \tag{2.5.13}$$

By the above lemma,

$$\max(0, \pi(\tilde{H})-n+r) \leq \pi(\tilde{H}_{11}) \leq \min(r, \pi(\tilde{H})) \tag{2.5.14}$$

and by Sylvester's law of inertia

$$\pi(\tilde{H}) = \pi(H) = \pi(B^*HB). \tag{2.5.15}$$

Since B^*AB and C^*AC are diagonal,

$$\pi(B^*HB) = \pi(B^*AB) \text{ and } \pi(C^*HC) = \pi(C^*AC). \tag{2.5.16}$$

From (2.5.13)-(2.5.16), the required result follows. ▲▲▲

COROLLARY 2.5.1. Let C be a matrix of order n x m with full row rank. If B^*AB and C^*AC are diagonal and B is invertible then $\pi(C^*AC) = \pi(B^*AB)$.

Proof. The proof is almost obvious. Since r=n, from (2.5.12), we have

$$\pi(B^*AB) \leq \pi(C^*AC) \leq \pi(B^*AB)$$

and the corollary follows. ▲▲▲

REMARK 2.5.1. Theorems 2.5.6 and 2.5.7 and Corollary 2.5.1 are valid if the matrices concerned are normal instead of diagonal.

REMARK 2.5.2. Inequalities analogous to (2.5.12) can be easily obtained for the number of eigenvalues in any open half plane by considering $Ae^{i\alpha}$ in the place of A.

REMARK 2.5.3. Applying Cauchy's inequalities mentioned earlier, a proof of $\pi(S^*HS) \leq \pi(H)$ for a Hermitian H and a rectangular matrix S is sketched by Hill in [60].

2.6. Generalization of Sylvester's Law of Inertia

Recently, Thompson [86] obtained the following elegant result consisting of four inequalities which involve the eigenvalues of two Hermitian matrices A and C^*AC and the singular values of the rectangular matrix C, i.e., the eigenvalues of the positive semidefinite matrix $(C^*C)^{1/2}$.

THEOREM 2.6.1 (Thompson [86]). Let $\alpha_1 \geq \ldots \geq \alpha_n$ be the eigenvalues of $A \in H_n$, $s_1 \geq \ldots \geq s_m$ be the singular values of $C \in M_{n,m}$ and $\beta_1 \geq \ldots \geq \beta_m$ be the eigenvalues of C^*AC. If $1 \leq i \leq m$, $1 \leq j \leq n$, $i+j-1 \leq m$, then

$$\beta_{i+j-1} \leq s_i^2 \alpha_j \quad \text{when } \alpha_j \geq 0, \tag{2.6.1}$$

$$\beta_{i+j-1} \leq s_{m+1-i}^2 \alpha_j \quad \text{when } \alpha_j \leq 0. \tag{2.6.2}$$

If $1 \leq i \leq m$, $1 \leq j \leq n$, $i+j > n$, then

$$\beta_{i+j-n} \geq s_i^2 \alpha_j \quad \text{when } \alpha_j \geq 0, \tag{2.6.3}$$

$$\beta_{i+j-n} \geq s_{m+1-i}^2 \alpha_j \quad \text{when } \alpha_j \leq 0. \tag{2.6.4}$$

These inequalities have a surprising amount of content, since they contain as special cases (i) Sylvester's law of

inertia (ii) its sharpened forms due to Ostrowski [69,70] and (iii) Cauchy's interlacing inequalities.

Considering m=n and C nonsingular let us put i=1 in (2.6.1)-(2.6.2) and i=n in (2.6.3)-(2.6.4). Then we arrive at the inequalities

$$s_n^2\alpha_j \leq \beta_j \leq s_1^2\alpha_j \text{ when } \alpha_j \geq 0 \qquad (2.6.5)$$

and

$$s_1^2\alpha_j \leq \beta_j \leq s_n^2\alpha_j \text{ when } \alpha_j \leq 0 \qquad (2.6.6)$$

or, equivalently

$$\beta_j = \phi_j\alpha_j , \quad 1 \leq j \leq n \qquad (2.6.7)$$

where

$$s_n^2 \leq \phi_j \leq s_1^2, 1 \leq j \leq n . \qquad (2.6.8)$$

In literature, the last mentioned result is known as Ostrowski's quantitative formulation of Sylvester's law of inertia. An immediate inference from this result is that if $A \in H_n$ and C is nonsingular then A and C^*AC have the same number of positive, negative, and vanishing eigenvalues, which, of course, is Sylvester's law of inertia.

Ostrowski's extension [70] of (2.6.5)-(2.6.6) for the case when C is n x m matrix is equivalent to

$$\beta_j \leq s_1^2\alpha_j \qquad \text{if } \beta_j > 0 \qquad (2.6.9)$$

and

$$\beta_{m+1-j} \geq s_1^2\alpha_{n+1-j} \text{ if } \beta_{m+1-j} < 0 \qquad (2.6.10)$$

and these two inequalities are also deducible [86] from (2.6.1)-(2.6.4).

Finally, assuming $m \leq n$ and taking $C = \begin{bmatrix} I_m \\ 0 \end{bmatrix}$, we see that the singular values of C are $s_1=\ldots=s_m=1$ and C^*AC is the leading principal submatrix of order m of A. With i=1, from (2.6.1)-(2.6.2) we have

$$\beta_j \leq \alpha_j \quad \text{for } j=1,\ldots,m. \tag{2.6.11}$$

Similarly by taking i=m and substituting j+n-m for j in (2.6.3)-(2.6.4) it follows that

$$\beta_j \geq \alpha_{j+n-m} \quad \text{for } j=1,\ldots,m. \tag{2.6.12}$$

Combination of (2.6.11) and (2.6.12) constitute the so-called Cauchy's interlacing inequalities.

All these deductions show the importance of Thompson's inequalities (2.6.1)-(2.6.4). In what follows we shall give generalization of these inequalities treating A and C^*AC as normal.

THEOREM 2.6.2. Let A be a normal matrix with eigenvalues $\{\alpha_j\}$ such that $\mathrm{Re}(\alpha_1) \geq \ldots \geq \mathrm{Re}(\alpha_n)$ and C be an n x m matrix with singular values $s_1 \geq \ldots \geq s_m$. Suppose C^*AC is normal having the eigenvalues $\{\beta_j\}$ such that $\mathrm{Re}(\beta_1) \geq \ldots \geq \mathrm{Re}(\beta_m)$.

If $1 \leq i \leq m$, $1 \leq j \leq n$, $i+j-1 \leq m$, then

$$\mathrm{Re}(\beta_{i+j-1}) \leq s_i^2 \mathrm{Re}(\alpha_j) \quad \text{when } \mathrm{Re}(\alpha_j) \geq 0, \tag{2.6.13}$$

$$\mathrm{Re}(\beta_{i+j-1}) \leq s_{m+1-i}^2 \mathrm{Re}(\alpha_j) \text{ when } \mathrm{Re}(\alpha_j) \leq 0. \tag{2.6.14}$$

If $1 \leq i \leq m$, $1 \leq j \leq n$, $i+j > n$, then

$$\mathrm{Re}(\beta_{i+j-n}) \geq s_i^2 \mathrm{Re}(\alpha_j) \quad \text{when } \mathrm{Re}(\alpha_j) \geq 0, \tag{2.6.15}$$

$$\mathrm{Re}(\beta_{i+j-n}) \geq s_{m+1-i}^2 \mathrm{Re}(\alpha_j) \text{ when } \mathrm{Re}(\alpha_j) \leq 0. \tag{2.6.16}$$

Proof. It is very simple. Since A is normal the eigenvalues of Re(A) i.e., $(A+A^*)/2$ are $Re(\alpha_j)$, $j=1,\ldots,n$. Similarly the eigenvalues of $C^*Re(A)C$, i.e., $Re(C^*AC)$ are $Re(\beta_j)$, $j=1,\ldots,m$. Applying Thompson's theorem with Re(A) instead of A yields (2.6.13)-(2.6.16). ΔΔΔ

As an immediate consequence of this result we have the following corollary which may be regarded as a sharpened form of Sylvester's law for normal matrices (Theorem 2.4.12).

COROLLARY 2.6.1. Let A and C^*AC be $n \times n$ normal matrices with C nonsingular. If the eigenvalues of A, C^*AC and the positive definite matrix $(C^*C)^{1/2}$ are respectively $\{\alpha_j\}$, $\{\beta_j\}$ and $\{s_j\}$ with $Re(\alpha_1) \geq \ldots \geq Re(\alpha_n)$, $Re(\beta_1) \geq \ldots \geq Re(\beta_n)$ and $s_1 \geq \ldots \geq s_n$, then

$$Re(\beta_j) = \phi_j Re(\alpha_j), \quad 1 \leq j \leq n \tag{2.6.17}$$

where

$$s_n^2 \leq \phi_j \leq s_1^2, 1 \leq j \leq n. \tag{2.6.18}$$

The above theorem and its corollary have similar analogues for the imaginary parts of the eigenvalues of A and C^*AC. For this we have to consider Im(A), i.e., $(A-A^*)/2i$. It may be noted that when the real parts of the eigenvalues are ordered decreasingly, then the corresponding imaginary parts need not be in the decreasing order. The ordering is considered separately for the real and imaginary parts.

We remark that if A is nonsingular and $\Theta[A] = \Theta[U]$, then A need not be normal in general. To show this we take $A = \begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix}$ so that $P = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ and $U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. It is easily verifiable that $\Theta[A] = \Theta[U]$. However A is not normal.

Next, let us consider the nonnormal matrix $A = \begin{bmatrix} 2 & i \\ 1 & i \end{bmatrix}$. For this, the polar factors are $P = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and $U = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$. Since the eigenvalues of A are $(2+i \pm \sqrt{3})/2$, obviously $\Theta[A] \neq \Theta[U]$.

The above two examples indicate that when A is nonsingular and nonnormal, A and its polar unitary factor may or may not be equiangular.

We have seen in Section 2.4 that when $P > 0$ and K is co-Hermitian then $\Theta[PK] = \Theta[K]$. Now it is interesting to inquire whether or not this result holds when K is normal in general. It is readily seen that the answer is in the negative, by taking P and K respectively as P and U of the second example given above. However, if PK is also assumed to be normal then it can be established that $\Theta[PK] = \Theta[K]$.

Before coming to this result, we shall say something about the product of two normal matrices. It is well known that if two normal matrices commute then their product is also a normal matrix [307]. But its converse is not true, since for $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, A, B and AB are normal whereas $AB \neq BA$.

Applying the result of Wiegmann [307] that if A and B are normal then AB (and hence BA) is normal iff each one of A and B commutes with the Hermitian polar factor of the other, Cullen [286] has provided a sufficient condition on a normal matrix A so that for normal B, we have AB normal iff A and B commute.

According to Cullen [286], $A \in M_n$ is said to have modularly distinct eigenvalues if unequal eigenvalues of A have unequal moduli.

THEOREM 2.7.2 (Cullen [286]). If A and B are normal and one of the two has modularly distinct eigenvalues, then AB is normal iff A and B commute.

One part of the theorem is obvious and for the other part we shall give a direct proof without applying Wiegmann's result stated earlier. Our proof depends on the following easy lemmas.

LEMMA 2.7.1. Let $(C_{ij})_{i,j=1,2}$ be the partition form of a normal matrix C where C_{11} and C_{22} are square matrices. Then $\|C_{21}\| = \|C_{12}\|$ where $\|X\|$ denotes the Frobenius norm $\{\mathrm{tr}(X^*X)\}^{1/2}$.

Proof. Since $C^*C = CC^*$, we have

$$C_{11}^*C_{11} + C_{21}^*C_{21} = C_{11}C_{11}^* + C_{12}C_{12}^*.$$

Taking trace for both sides, the result follows. ▲▲▲

COROLLARY 2.7.1. Let $C = \begin{bmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{bmatrix}$, where C_{11} and C_{22} are square matrices. Then C is normal iff $C_{21} = 0$ and C_{11}, C_{22} are normal.

Proof. It is an immediate consequence of the above lemma. ΔΔΔ

LEMMA 2.7.2. Let $D = \text{diag}(d_1,\ldots,d_n)$ have modularly distinct eigenvalues with $|d_1| \leq \ldots \leq |d_n|$ and let C be a normal matrix. Then if either CD or DC is normal then C and D commute.

Proof. Let us assume that DC is normal with $C = (c_{ij})$. There may be two cases. First let us consider the case $|d_1| < |d_2|$. Partitioning C and DC with 1 x 1 block in the leading position for both matrices, an application of the preceding lemma gives

$$\sum_{i=2}^{n} |c_{i1}|^2 = \sum_{j=2}^{n} |c_{1j}|^2 \tag{2.7.1}$$

and

$$\sum_{i=2}^{n} |d_i c_{i1}|^2 = \sum_{j=2}^{n} |d_1 c_{1j}|^2 \tag{2.7.2}$$

From these two relations, it follows that

$$\sum_{i=2}^{n} \{(|d_i|^2 - |d_1|^2)|c_{i1}|^2\} = 0. \tag{2.7.3}$$

By the assumption that $|d_1| < |d_2|$, we have $|d_1| < |d_i|$ for $i=2,\ldots,n$ and hence it follows that $c_{i1}=0$ for $j=2,\ldots,n$. Hence proving DC = CD reduces to a similar problem in dimension n-1.

Next, we shall consider the possibility that

$$|d_1| = |d_2| = \ldots = |d_m| < |d_{m+1}| \tag{2.7.4}$$

where $m \leq n$. Of course, when m=n, there is no inequality term in (2.7.4). Since D has modularly distinct eigenvalues, (2.7.4) implies that $d_1=d_2=\ldots=d_m=d$, say. Let us write $D = \mathrm{diag}(dI_m, \tilde{D})$ and $C = (C_{ij})_{i,j=1,2}$ in partitioned forms where $C_{11} \in M_m$. Now,

$$DC = \begin{bmatrix} dC_{11} & dC_{12} \\ \tilde{D}C_{21} & \tilde{D}C_{22} \end{bmatrix}.$$

Since C and DC are normal, evidently

$$\|C_{21}\| = \|C_{12}\| \tag{2.7.5}$$

and

$$\|\tilde{D}C_{21}\| = \|dC_{12}\|. \tag{2.7.6}$$

Hence

$$\|\tilde{D}C_{21}\| = |d|\,\|C_{21}\| \tag{2.7.7}$$

which may be expressed in the form

$$\sum_{j=1}^{m} \sum_{i=m+1}^{n} \{(|d_i|^2-|d|^2)\,|c_{ij}|^2\} = 0. \tag{2.7.8}$$

By the hypothesis, $|d_i| > |d|$ for i=m+1,...,n. Thus (2.7.8) yields that $C_{21} = 0$ and hence we have $C_{12} = 0$. Again we see that the problem of showing CD = DC reduces to a similar problem of lower dimension and in this case it is of dimension n-m.

Since n is finite and the lemma is true for the set of 1x1 matrices, the result follows. In the same manner one

can proceed to prove the result if CD is assumed to be normal. ▲▲▲

Now we shall complete the proof of Theorem 2.7.2. Let us assume that A has modularly distinct eigenvalues. Then there exists a unitary matrix V such that $V^*AV = D = \text{diag}(d_1,\ldots,d_n)$ with entries as described in the above lemma. Denoting V^*BV by C, AB becomes $VDCV^*$. Now C and DC are normal. Applying the preceding lemma, it follows that CD = DC and therefore, AB = BA. The same argument remains valid if we interchange the role of A and B. The proof of Theorem 2.7.2 is now complete. ▲▲▲

Since positive semidefinite matrices always have modularly distinct eigenvalues, as an application of Theorem 2.7.2, we have the following theorem.

THEOREM 2.7.3. Let $P \geq 0$ and K be normal such that PK is also normal. Then

(i) $\Theta[PK] = \Theta[K]$ if $P > 0$ and

(ii) $\Theta[PK] \leq \Theta[K]$ if $P \geq 0$.

Proof. By Cullen's theorem PK = KP and hence the argument given in Theorem 2.7.1 is applicable and the theorem follows at once. ▲▲▲

We remark that this theorem seems to be a generalization of Theorem 2.7.1.

2.8. Totally Normal Matrices

In this section and the following one, we shall deal with normal matrices having some normal principal submatrices.

In analogy with positive definite matrices, i.e., matrices all whose leading principal submatrices are again positive definite (in fact, all whose principal submatrices are positive definite) and with totally nonnegative (positive) matrices, i.e., matrices whose minors of all orders are nonnegative (positive)[8, p.118], we define totally normal matrices as given below. The notion would allow us to make a generalization of Cauchy's interlacing inequalities.

DEFINITION 2.8.1. $A \in M_n$ is said to be totally normal if every principal submatrix of A is normal.

It is clear that totally normal matrices are normal and not conversely. Obvious examples of totally normal matrices are diagonal matrices and co-Hermitian matrices.

Listed below are some observations about totally normal matrices.

(1) If $A \in M_n$ and $B \in M_m$ are totally normal matrices, then their direct sum is also totally normal.

(2) We know that if $A \in M_n$ and $B \in M_m$ are normal, then their Kronecker product $A \otimes B$ is normal [16, p.70]. However, similar result is not true for totally normal matrices. For example, $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ and $B \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix}$ are totally normal, but $A \otimes B$ is not totally normal.

(3) Another well-known fact about normal matrices is that the sum as well as the product of two commuting normal matrices are again normal. This type of result is not true for totally normal matrices. To see this, consider the totally normal matrices

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \end{bmatrix}.$$

In view of the circulant nature of A and B here, they commute. Simple calculations reveal that neither A+B nor AB is totally normal.

The construction of counterexamples of this type provides a justification for the study of certain special classes of matrices (e.g. circulants).

(4) If A is totally normal then A-zI is also totally normal where z is any complex number. Hence, by virtue of Lemma 2.4.2, we conclude that a normal matrix having all its eigenvalues along a straight line is totally normal.

(5) If A is totally normal and U is unitary, then U^*AU need not be totally normal in general. However, if U happens to be a permutation matrix then U^*AU is always totally normal.

(6) If A is totally normal and of rank r, then there exists a permutation matrix P such that P^*AP has a nonsingular normal matrix of order r in the top left corner. To prove this it suffices to show that there exists a nonsingular

principal submatrix of order r of A. Suppose if all the r-th order principal submatrices are singular, then the characteristic polynomial reduces to

$$x^n - s_1 x^{n-1} + \ldots + (-1)^{r-1} s_{r-1} x^{n-r+1}$$

where s_k denotes the sum of all principal minors of order k of A [13, p.54]. This shows that A has at least (n-r+1) vanishing eigenvalues which is a contradiction since any normal matrix of rank r has precisely (n-r) zero eigenvalues.

In what follows first we shall give a characterization of totally normal matrices. As a prelude, we study the 2 x 2 and 3 x 3 situations and subsequently we take up the general case.

LEMMA 2.8.1. $A = (a_{rs})_{r,s=1,2}$ is totally normal iff

$$|a_{12}| = |a_{21}| \tag{2.8.1}$$

and

$$\bar{a}_{21} d_{12} = a_{12} \bar{d}_{12} \tag{2.8.2}$$

where

$$d_{12} = a_{11} - a_{22}. \tag{2.8.3}$$

Proof. Note that a 2 x 2 matrix is totally normal iff it is normal. Now, by equating the corresponding elements in AA^* and A^*A the result follows. ▲▲▲

Here is another form of this lemma:

$A = (a_{rs})_{r,s=1,2}$ is totally normal iff

$$a_{12} = \bar{a}_{21} \exp(i\phi) \tag{2.8.4}$$

and

$$d_{12} = \bar{d}_{12} \exp(i\phi) \tag{2.8.5}$$

for some real ϕ with d_{12} defined as in (2.8.3).

In the following lemma, n=3. However, for future reference, the lemma is framed in terms of n.

LEMMA 2.8.2. $A = (a_{rs}) \in M_n$ (with n=3) is totally normal iff for every r and s satisfying $1 \leq r < s \leq n$,

$$|a_{rs}| = |a_{sr}|, \tag{2.8.6}$$

$$\bar{a}_{sr}d_{rs} = a_{rs}\bar{d}_{rs} \tag{2.8.7}$$

and

$$a_{rk}\bar{a}_{sk} = \bar{a}_{kr}a_{ks} \text{ for all } k \in \{1,2,\ldots,n\}\backslash\{r,s\} \tag{2.8.8}$$

where

$$d_{rs} = a_{rr} - a_{ss}. \tag{2.8.9}$$

Proof. In view of the lemma proved just now, it is evident that the conditions given in (2.8.6) and (2.8.7) are necessary and sufficient to ensure the normality of all the three second order principal submatrices of A.

Now by equating the entries in (1,2) position of AA^* and A^*A we have

$$\sum_{k=1}^{3} a_{1k}\bar{a}_{2k} = \sum_{k=1}^{3} \bar{a}_{k1}a_{k2}. \tag{2.8.10}$$

By subtracting the equation (2.8.7) with r=1, s=2 from (2.8.10), we find

$$a_{13}\bar{a}_{23} = \bar{a}_{31}a_{32}. \tag{2.8.11}$$

Similarly by considering the elements in (1,3) and (2,3) positions in the products AA^* and A^*A and by making use of (2.8.7) with suitable r and s, the other two conditions in (2.8.8) follow. This completes the proof. ΔΔΔ

We now turn to the characterization of $n \times n$ ($n \geq 3$) totally normal matrices.

THEOREM 2.8.1. $A \in M_n$, $n \geq 3$ is totally normal iff all its third order principal submatrices are totally normal, or equivalently, iff all its second and third order principal submatrices are normal.

Proof. One part is obvious. To prove the other part let us assume that all the third order principal submatrices of A are totally normal. By applying the preceding lemma to each one of the third order principal submatrices, we see that for every r and s satisfying $1 \leq r < s \leq n$, (2.8.6)-(2.8.9) hold.

Following Marcus and Minc [16, Chapter I, Section 2], let $Q_{m,n}$, $1 \leq m \leq n$ denote the set of all strictly increasing sequences of m integers chosen from $1,2,\ldots,n$. If $\alpha = (i_1,\ldots,i_m) \in Q_{m,n}$, we designate the principal submatrix of A lying in rows α and columns α with the notation $A[\alpha|\alpha]$. We are through if we show that $A[\alpha|\alpha]$ is normal for every $\alpha \in Q_{m,n}$, $m=1,\ldots,n$. We shall, therefore, consider one such $A[\alpha|\alpha]$ for an arbitrary $m \in \{1,\ldots,n\}$ and denote it by B. If $B = (b_{rs})$ $(r,s=1,\ldots,m)$ then $b_{rs} = a_{i_r i_s}$. Again by using the formula $BB^* = B^*B$ for B to be normal and taking advantage of the Hermitian nature of these products, we see the necessary and sufficient conditions for B to be normal as

$$\sum_{k=1}^{m} b_{rk}\bar{b}_{sk} = \sum_{k=1}^{m} \bar{b}_{kr}b_{ks}, \qquad 1 \leq r \leq s \leq m. \tag{2.8.12}$$

The above relation for the case r=s reduces to

$$\sum_{k=1}^{m} (|b_{sk}|^2 - |b_{ks}|^2) = 0. \tag{2.8.13}$$

In case $r \neq s$, (2.8.12) can be rewritten as

$$\sum_{\substack{k=1 \\ k\neq r,s}}^{m} (b_{rk}\bar{b}_{sk} - \bar{b}_{kr}b_{ks}) + \bar{b}_{sr}(b_{rr}-b_{ss}) - b_{rs}(\bar{b}_{rr}-\bar{b}_{ss}) = 0. \tag{2.8.14}$$

We know that $b_{rs} = a_{i_r i_s}$. Moreover, $i_r < i_s$ whenever $r < s$. Hence the relations (2.8.13) and (2.8.14) are certainly valid in view of (2.8.6)-(2.8.9). Thus, B is normal. This completes the proof of the theorem. ▲▲▲

From the proof, we observe that an equivalent form of the above theorem is Lemma 2.8.2 with $n \geq 3$.

COROLLARY 2.8.1. For any totally normal matrix $A = (a_{rs})$ having no zero off-diagonal element, there exists a real ϕ such that

$$a_{rs} = \bar{a}_{sr}\exp(i\phi), \quad 1 \leq r < s \leq n \tag{2.8.15}$$

and

$$d_{rs} = \bar{d}_{rs}\exp(i\phi), \quad 1 \leq r < s \leq n \tag{2.8.16}$$

where d_{rs} is defined as in (2.8.9).

Proof. In view of (2.8.4)-(2.8.5), the conditions (2.8.6)-(2.8.7) can be rephrased as

$$a_{rs} = \bar{a}_{sr}\exp(i\phi_{rs}) \tag{2.8.17}$$

and

$$d_{rs} = \bar{d}_{rs}\exp(i\phi_{rs}) \tag{2.8.18}$$

for some real ϕ_{rs}, $1 \leq r < s \leq n$. By considering $r < s < k$,

let us substitute for a_{rk} and a_{sk} in (2.8.8) by using (2.8.17). This yields

$$\bar{a}_{kr}\exp(i\phi_{rk})a_{ks}\exp(-i\phi_{sk}) = \bar{a}_{kr}a_{ks}. \qquad (2.8.19)$$

Since the off-diagonal elements are assumed to be nonzero, it follows that $\phi_{rk} = \phi_{sk}$ for $r < s < k$, with the convention that for all r and s, $0 \leq \phi_{rs} < 2\pi$. By a similar procedure, it can be easily proved that $\phi_{kr} = \phi_{ks}$ for $k < r < s$. Hence we have $\phi_{rs} = \phi$ say, for all r and s satisfying $1 \leq r < s \leq n$. The proof of the corollary is now complete. ▲▲▲

REMARK 2.8.1. Any matrix $A = (a_{rs})$ satisfying (2.8.15) and (2.8.16) is of course totally normal even if some off-diagonal elements are zero, because these conditions meet the sufficient requirements (2.8.6)-(2.8.9) for a matrix to be totally normal.

If we assume that only all the leading principal submatrices of A are given to be normal, it can be easily seen by routine manipulations that (2.8.6) holds for all r and s; also for certain values of r, s, k, (2.8.8) holds. This motivates to know whether the normality of all leading principal submatrices will suffice to say that A is totally normal. It cannot be so, however, since in

$$\begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ -1 & -1 & 0 \end{bmatrix}$$

all the three leading principal submatrices are normal but

the principal submatrix $\begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix}$ is not normal.

We conclude this section by stating Cauchy's interlacing inequalities for principal submatrices of totally normal matrices.

THEOREM 2.8.2. Let B be any principal submatrix of order m of a totally normal matrix $A \in M_n$. If the eigenvalues of A and B are respectively $\{\alpha_j\}$ and $\{\beta_j\}$ such that $\mathrm{Re}(\alpha_1) \geq \ldots \geq \mathrm{Re}(\alpha_n)$ and $\mathrm{Re}(\beta_1) \geq \ldots \geq \mathrm{Re}(\beta_m)$ then

$$\mathrm{Re}(\alpha_j) \geq \mathrm{Re}(\beta_j) \geq \mathrm{Re}(\alpha_{j+n-m}), \quad j=1,\ldots,m. \qquad (2.8.20)$$

Proof. It follows by applying the same principle used in Theorem 2.6.2. ▲▲▲

In fact, for the inequalities in (2.8.20) to hold, it is enough that A and B are normal.

2.9. Angularity and Inertia of a Partitioned Normal Matrix

We have already seen in the survey chapter that if $(H_{ij})_{i,j=1,2}$ is the partitioned form of $H \in H_n$ where H_{11} is nonsingular, then

$$\mathrm{In}(H) = \mathrm{In}(H_{11}) + \mathrm{In}(K_{22}),$$

K_{22} being the Schur complement of H_{11} in H given by

$$K_{22} = H_{22} - H_{12}^{*}H_{11}^{-1}H_{12}.$$

This interesting inertia result is due to Haynsworth [57].

In [1, p.97], it has been remarked by Barnett that no result of this nature is known for non-Hermitian matrices.

In this section, as an application of our basic angularity theorem (Theorem 2.4.1), we prove some results for angularity and inertia of partitioned normal matrices, when the block matrices fulfil certain conditions.

To obtain our main result of this section, we require the following lemma.

LEMMA 2.9.1. Let A be an $n \times n$ normal matrix partitioned in the form $(A_{ij})_{i,j=1,2}$ where A_{11} and A_{22} are normal, A_{11} is nonsingular and

$$A_{11}^* A_{12} = A_{11} A_{21}^*. \tag{2.9.1}$$

Then the Schur complement B_{22} of A_{11} in A defined by

$$B_{22} = A_{22} - A_{21} A_{11}^{-1} A_{12}$$

is also normal.

Proof. Since A_{11} and A_{22} are normal we have

$$A_{11} A_{11}^* = A_{11}^* A_{11} \tag{2.9.2}$$

and

$$A_{22} A_{22}^* = A_{22}^* A_{22}. \tag{2.9.3}$$

In view of (2.9.1)-(2.9.3), the normality of A yields

$$A_{12} A_{12}^* = A_{21}^* A_{21} \tag{2.9.4}$$

$$A_{21} A_{21}^* = A_{12}^* A_{12} \tag{2.9.5}$$

and

$$A_{12} A_{22}^* = A_{21}^* A_{22}. \tag{2.9.6}$$

Also, from (2.9.1) and (2.9.2) we have

$$A_{11}^{-1} A_{12} = A_{11}^{-*} A_{21}^*. \tag{2.9.7}$$

Now, by making use of (2.9.3), (2.9.4), (2.9.6) and (2.9.7) it can be easily verified that

$$B_{22}B_{22}^{*} = B_{22}^{*}B_{22}.$$

This completes the proof of the lemma. ▲▲▲

REMARK 2.9.1. In order that B_{22} be normal, the assumption (2.9.1) cannot be dropped in general, as may be seen by an example. Consider

$$A = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & 1 \\ -1 & -1 & 1 \end{bmatrix} \text{ with } A_{11} = [1].$$

Then

$$B_{22} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} - \begin{bmatrix} -1 \\ -1 \end{bmatrix}[1][1 \ 1] = \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix}$$

which is obviously not normal.

We can now prove the principal result of this section.

THEOREM 2.9.1. Under the hypotheses of Lemma 2.9.1,

$$\theta[A]_{\omega} = \theta[A_{11}]_{\omega} + \theta[B_{22}]_{\omega} \text{ for all } \omega \in \Omega.$$

Proof. The proof is almost along the same lines as in Theorem 1 of Haynsworth [57]. Let

$$C = \begin{bmatrix} I_m & -A_{11}^{-1}A_{12} \\ 0 & I_{n-m} \end{bmatrix}$$

m being the order of A_{11}. By direct computations we have

$$C^{*}AC = \begin{bmatrix} A_{11} & 0 \\ L & B_{22} \end{bmatrix}$$

where $L = A_{21} - A_{12}^{*}A_{11}^{-*}A_{11}$. Because of (2.9.7), L vanishes. Hence, by means of the preceding lemma, $C^{*}AC$ becomes normal. Now by applying Corollary 2.4.3, the angularity result stated in the theorem follows at once. ▲▲▲

As an immediate consequence of this theorem we have

COROLLARY 2.9.1. Under the hypotheses of Lemma 2.9.1,

$$In(A) = In(A_{11}) + In(B_{22}).$$

The last mentioned formula shows that the inertia of a normal matrix can be determined as the sum of the inertias of two lower order normal matrices. Although this formula is applicable only to a certain type of normal matrices, it is of course, a generalization of Haynsworth's result.

COROLLARY 2.9.2. If $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix}$ is normal where A_{11} is nonsingular and normal, then

$$In(A) = In(A_{11}) + In(-A_{21}A_{11}^{-1}A_{12}).$$

Proof. Since the normality of the given form of A includes the condition $A_{11}^{*}A_{12} = A_{11}A_{21}^{*}$, this corollary follows from the previous one. ▲▲▲

COROLLARY 2.9.3. If A is as in the preceding corollary with an additional assumption that $A_{21} = A_{12}^{*}$ then

$$\pi(A) \leq m - \delta(A_{11})$$

$$\nu(A) \leq m - \delta(A_{11})$$

where m is the order of A_{11}.

Proof. Note that, by Lemma 2.9.1, $-A_{21}A_{11}^{-1}A_{12}$ i.e., $A_{12}^{*}(-A_{11}^{-1})A_{12}$ is normal. It is given that $-A_{11}^{-1}$ is normal. Hence, in view of Remark 2.5.1, we shall apply Theorem 2.5.6, so that

$$\begin{aligned}\pi(-A_{12}^{*}A_{11}^{-1}A_{12}) &\leq \pi(-A_{11}^{-1}) \\ &= \pi(-A_{11}) \\ &= \nu(A_{11}).\end{aligned}$$

Therefore, from the formula given in Corollary 2.9.2, we have

$$\begin{aligned}\pi(A) &\leq \pi(A_{11}) + \nu(A_{11}) \\ &= m - \delta(A_{11}).\end{aligned}$$

Similarly we can prove the other part. ΔΔΔ

We shall now present analogous statements of several results in [57], for partitioned normal matrices. By appealing to the familiar formula that $In(A+A^{*}) = In(A)$ for a normal A, these results can be easily proved. In the following theorems and corollaries of this section, by $In(A) \geq (a,b,c)$ we shall mean $\pi(A) \geq a$, $\nu(A) \geq b$ and $\delta(A) \geq c$ as in [57].

THEOREM 2.9.2. If $(A_{ij})_{i,j=1,2}$ is the partitioned form of a normal matrix A with A_{11} normal and $In(A_{11}) = (p,q,0)$ then $In(A) \geq (p,q,0)$.

THEOREM 2.9.3. If A_k and A_m are two normal principal submatrices of orders k and m respectively, of a normal matrix A and

$$In(A_k) = (p_k, q_k, 0)$$
$$In(A_m) = (p_m, q_m, 0)$$

then

$$In(A) \geq (p, q, 0)$$

where

$$p = \max(p_k, p_m) \text{ and } q = \max(q_k, q_m).$$

COROLLARY 2.9.4. If a normal matrix A has a positive stable normal principal submatrix of order p and a negative stable normal principal submatrix of order q, then $In(A) \geq (p, q, 0)$.

COROLLARY 2.9.5. If a normal matrix has a diagonal element with positive (negative) real part, then it has at least one eigenvalue with positive (negative) real part.

COROLLARY 2.9.6. If A is an $n \times n$ normal matrix in the partitioned form $(A_{ij})_{i,j=1,2}$ where A_{11} and A_{22} are respectively positive and negative stable normal matrices, then $In(A) = (m, n-m, 0)$, m being the order of A_{11}.

For all the results derived in this section, the matrix A and some of its principal submatrices are assumed to be normal. Hence these results are well applicable to totally normal matrices provided the block matrices meet other requirements stated in the theorems and corollaries.

3. LINEAR TRANSFORMATIONS WITH INVARIANTS

3.1. Introduction

Many problems in applied mathematics involve the study of transformations, that is the way in which certain input data is transformed into an output data. In many situations, these transformations are linear and moreover linear algebra is essentially the study of linear transformations.

Another key concept in algebra is the notion of invariance. If T is a linear transformation of V into itself then a function f defined on V is said to be invariant under the operator T if

$$f(T(x)) = f(x) \text{ for every } x \in V. \qquad (3.1.1)$$

In certain situations one might wish to simplify an original problem through a transformation but with a restriction that a given character of elements in V remain unchanged in the process; while in other situations one may wish to infer some property of x from that of Tx which may be available as a data. Furthermore, while analysing various mathematical problems one often comes across relations of type (3.1.1). In all such cases structure theorems for T turn out to be very useful. This motivates the study of determining the structure of all linear operators on M_n, $\mathbb{R}^n$, $\mathbb{C}^n$ etc. which map certain subsets into themselves and at the same time leave certain quantities invariant.

In Section 3.2 of this chapter we characterize the structure of linear operators on M_n having inertia and angularity as invariants over the classes of Hermitian and normal matrices. Sections 3.3 and 3.4 are devoted to the study of matrix transformations which preserve number of sign changes (variations) and nondecreasing trend of vectors in $\mathbb{R}^n$. In Section 3.5, we deal with linear operators on $\mathbb{C}^n$ which preserve inertia, angularity and number of zero components. Using some of the results of Section 3.5, in Section 3.6 we determine the structure of linear transformations which map circulants into circulants and preserve their inertia and angularity.

3.2. Inertia- and Angularity-preserving Transformations on H_n and N_n.

In this section we consider the following three problems of characterizing all linear transformations T of M_n into M_n such that

(a) $T(H_n) \subseteq H_n$ and $In(T(H)) = In(H)$ for all $H \in H_n$,

(b) $T(N_n) \subseteq N_n$ and $\theta[T(N)] = \theta[N]$ for all $N \in N_n$, and

(c) $T(N_n) \subseteq N_n$ and $In(T(N)) = In(N)$ for all $N \in N_n$

where H_n and N_n respectively denote the classes of all Hermitian and normal matrices in M_n.

As will be seen these characterizations are interdependent at least in the sense that the proof in the case (c) depends on the characterization in the case (b) and that the proof

in the case (b) depends on the characterization in the case (a). Another interesting fact which emerges is that the operators T have the same structure in the latter two problems.

THEOREM 3.2.1. Let $T: M_n \to M_n$ be a linear transformation. Then, for all $H \in H_n$, we have $T(H) \in H_n$ and $In(T(H)) = In(H)$ iff there exists a fixed nonsingular matrix C such that

$$T(A) = C^*AC \text{ for all } A \in M_n \tag{3.2.1}$$

or

$$T(A) = C^*A'C \text{ for all } A \in M_n . \tag{3.2.2}$$

Proof. The sufficiency follows from Sylvester's theorem. For the necessity, note that T(I) must be positive definite, I denoting the identity matrix. Let Q denote the positive definite square root of $(T(I))^{-1}$, and set $S(A) = Q^*T(A)Q$ for every $A \in M_n$. Then, if K is Hermitian, so is S(K), and, by Sylvester's theorem $In(S(K)) = In(K)$. In particular, if H is Hermitian,

$$\delta(H-\lambda I) = \delta(S(H-\lambda I)) = \delta(S(H)-\lambda I) \text{ for all real } \lambda \tag{3.2.3}$$

because $H-\lambda I$ is Hermitian and $S(I) = I$. But all the eigenvalues of H and S(H) are real; so (3.2.3) implies that H and S(H) have the same eigenvalues with the same multiplicities. Now Theorem 1.3.4 of Marcus and Moyls becomes applicable, and therefore

$$S(A) = U^*AU \text{ for all } A \in M_n$$

or

$$S(A) = U^*A'U \text{ for all } A \in M_n$$

for some unitary U. Setting $C = UQ^{-1}$, Theorem 3.2.1 follows. $\Delta\Delta\Delta$

Let us now turn to a discussion of linear transformations T on M_n mapping normal matrices into themselves and preserving inertia for each normal matrix. In the problem of directly determining the structure of inertia-preserving transformation T, we face some difficulties whereas in the corresponding problem of angularity-preserving transformations there is no such difficulty. At the same time we are able to determine the structure of inertia-preserving transformations from that of angularity-preserving transformations, thus justifying our remark in the beginning of the previous chapter that working with angularity will be easier than with inertia in certain situations. Therefore, we shall first characterize the angularity-preserving linear transformations on N_n.

THEOREM 3.2.2. Let $T: M_n \to M_n$ be a linear transformation. Then for all $N \in N_n$, we have $T(N) \in N_n$ and $\theta[T(N)] = \theta[N]$ iff there exists a fixed matrix C such that C is a nonzero scalar multiple of a unitary matrix and T has the form (3.2.1) or (3.2.2).

Proof. The sufficiency is obvious. For, if $T(A) = kU^*AU$ for all $A \in M_n$ or $T(A) = kU^*A'U$ for all $A \in M_n$ with k positive and U unitary, then T(N) is normal for every normal N. Moreover, by Corollary 2.4.3, $\theta[T(N)] = \theta[N]$ for all $N \in N_n$.

Next, for the necessity, let H be Hermitian, and note

that T(H) must be a normal matrix with real spectrum, that is, Hermitian. Furthermore $In(T(H)) = In(H)$, and so, by Theorem 3.2.1, there exists a nonsingular matrix C such that T either maps all $A \in M_n$ into C^*AC or into $C^*A'C$. Theorem 2.4.5 now shows that C is a nonzero multiple of a unitary matrix. ▲▲▲

In order to deduce the structure of inertia-preserving linear transformations on N_n from the above theorem we shall prove the following theorem.

THEOREM 3.2.3. Let $T: M_n \to M_n$ be a linear transformation. Then T preserves inertia for each $N \in N_n$ iff it preserves angularity for each $N \in N_n$.

Proof. If T preserves angularity evidently it preserves inertia. To prove the converse let us assume that $In(T(N)) = In(N)$ for all $N \in N_n$. Consider any real number α. For $N \in N_n$, $e^{i\alpha}N \in N_n$, and hence by making use of the linearity of T, $In(e^{i\alpha}T(N)) = In(T(e^{i\alpha}N)) = In(e^{i\alpha}N)$ which in turn implies that $\pi(e^{i\alpha}T(N)) = \pi(e^{i\alpha}N)$ for any arbitrary real α. This is equivalent to saying that T(N) and N have equal number of eigenvalues in any arbitrary open half plane. Now applying the same technique used in the proof of Theorem 2.4.1, it follows that $\theta[T(N)] = \theta[N]$ for all $N \in N_n$. This completes the proof of the theorem. ▲▲▲

REMARK 3.2.1. In fact, the theorem just proved is true even if N_n is replaced by any other class S_n in M_n having the property that S_n is closed under scalar multiplication.

As an immediate consequence of Theorems 3.2.2 and 3.2.3 we have

THEOREM 3.2.4. Let $T: M_n \to M_n$ be a linear transformation. Then for all $N \in N_n$ we have $T(N) \in N_n$ and $In(T(N)) = In(N)$ iff there exists a fixed matrix C such that $C = kU$ where $k \neq 0$ and U is a unitary matrix and T has the form (3.2.1) or (3.2.2).

3.3. Structure of VP_0 and VP_2 Matrices

The concept of variation-diminishing transformations was introduced by Schoenberg in 1930, an exhaustive account of which, including a wide range of applications is available in Karlin [11]. Gantmacher and Krein [9] elaborated rather completely the various characteristic forms of matrix variation-diminishing transformations. Some particularly interesting papers in this connection are due to Gantmacher and Krein, Karlin and MacGregor, and Pólya and Schoenberg as cited in Bellman [4, p.314]. In this context, the structure of variation-preserving linear operator $L: C[0,1] \to C[0,1]$, $C[0,1]$ denoting the space of all continuous real valued functions on the closed interval $[0,1]$, has recently been determined by Rathore [300]. In this section and the following one, we propose to study the structure of matrix variation-preserving transformations $A: \mathbb{R}^n \to \mathbb{R}^n$. As will be seen in the next section, the results in the discrete case involve some surprises as compared to the continuous case.

To begin our study, we shall recall certain well-known definitions. $A = (a_{ij}) \in M_n(\mathbb{R})$ is called nonnegative if $a_{ij} \geq 0$ for $i,j=1,\ldots,n$ and we write $A \geq 0$. Similarly a vector $x = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$ is said to be nonnegative if $x_i \geq 0$ for $i=1,\ldots,n$ and we write $x \geq 0$. Similarly we may conceive of the relations $A \leq 0$, $x \leq 0$, $x > 0$ etc. For $x, y \in \mathbb{R}^n$, the relation $x \geq y$ is equivalent to the statement $x-y \geq 0$.

Throughout the present section and the next, we use $A \geq 0$ to denote nonnegative matrix A (and not positive semidefinite matrix A as used elsewhere). Moreover all the matrices and vectors encountered in these two sections are assumed to be in $M_n(\mathbb{R})$ and $\mathbb{R}^n$ respectively. As usual, e_i $(i=1,\ldots,n)$ will denote the vector whose i-th coordinate is 1 and whose other coordinates are all zero. We shall represent the vector having all coordinates as 1 by e. Following Marcus and Minc [16], we denote the i-th row of A by $A_{(i)}$ and the j-th column of A by $A^{(j)}$ and it may be noted that $Ae_j = A^{(j)}$. By $(Ax)_i$ we mean the i-th coordinate of the vector Ax. If $A \geq 0$ and $Ae=e$, that is, every row sum of A is 1, then A is said to be row stochastic.

Here are some more definitions closely related to the matrix variation-preserving transformations.

DEFINITION 3.3.1. The variation $v(x)$ of $x = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$ is defined as the total number of sign changes in the

sequence $\{x_1,\ldots,x_n\}$ discarding the zero components in the sequence.

Thus for example, if $x = (0,-1,9,4,0,-5,0)^T$, then $v(x)=2$. Note that $v(x)=0$ iff $x \geq 0$ or $x \leq 0$.

DEFINITION 3.3.2. $A \in M_n(\mathbb{R})$ is said to be variation-preserving of order m (abbreviated VP_m), m being a nonnegative integer less than n, if $v(x) \leq m$ for $x \in \mathbb{R}^n$ always implies

$$v(Ax) = v(x). \tag{3.3.1}$$

If (3.3.1) holds for all $x \in \mathbb{R}^n$, then A is said to be variation-preserving (VP).

Obviously VP_m implies VP_{m-1} for m=1,2,...,n-1. Of course, VP and VP_{n-1} are one and the same.

DEFINITION 3.3.3. $x = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$ is said to be nondecreasing if $x_1 \leq \ldots \leq x_n$ and we write x is $\uparrow$. It is nonincreasing if $x_1 \geq \ldots \geq x_n$ and we write x is $\downarrow$. By a monotonic vector we mean either nondecreasing or nonincreasing vector.

DEFINITION 3.3.4. $A \in M_n(\mathbb{R})$ is said to be monotone-preserving (MP) if Ax is monotonic for all monotonic $x \in \mathbb{R}^n$.

DEFINITION 3.3.5. $A \in M_n(\mathbb{R})$ is said to be trend-preserving (TRP) if x is $\uparrow$ implies Ax is $\uparrow$ (or equivalently x is $\downarrow$ implies Ax is $\downarrow$). It is said to be

trend-reversing (TRR) if x is $\uparrow$ implies Ax is $\downarrow$ (or equivalently x is $\downarrow$ implies Ax is $\uparrow$). A is said to be consistently monotone-preserving (CMP) if it is either trend-preserving or trend-reversing.

Now we shall state some of the results obtained by Rathore [300] pertaining to variation-preserving operators in the continuous case, which are relevant to our study of such transformations in the discrete case. The definitions of v(f), that is, the variation of $f \in C[0,1]$, VP_m and VP are similar to those in the discrete case. However, in the continuous case n does not come into the picture.

THEOREM 3.3.1 [300, Theorem 1]. Let $L: C[0,1] \to C[0,1]$ be linear. Then L is VP_0 iff L is positive or negative or of the form

$$L(f;x) = \phi(f)\psi(x), \quad f \in C[0,1], \quad x \in [0,1] \qquad (3.3.2)$$

where ϕ is a linear functional on C[0,1] and $\psi \in C[0,1]$ is a nonnegative function.

THEOREM 3.3.2 [300, Theorem 2]. For a linear operator $L: C[0,1] \to C[0,1]$, the following statements are equivalent.

(a) L is VP.

(b) L is VP_2.

(c) For some η, $\zeta \in C[0,1]$ and satisfying (i) $v(\eta)=0$ (ii) ζ is monotone onto [0,1] and (iii) η vanishes on an interval $[\alpha,\beta] \subset [0,1]$ only if ζ is constant on $[\alpha,\beta]$, there holds

$$L(f;x) = n(x)f(\zeta(x)) \qquad (3.3.3)$$

for all $x \in [0,1]$ and $f \in C[0,1]$.

The object of this section is to prove the analogues of the above two theorems in the discrete case.

It has been remarked in [300] that in the continuous case that VP_m <==> VP for $m \geq 2$. Furthermore, it is shown that m=2 is the least such integer, by providing an example of a linear operator which is VP_1 but not VP_2. The validity of such a remark in the case of $A: \mathbb{R}^n \to \mathbb{R}^n$ will be investigated in the next section.

In the following theorem we characterize VP_o matrix transformations. This result serves as a basic tool in the proof of other matrix variation-preserving theorems.

THEOREM 3.3.3. $A = (a_{ij}) \in M_n(\mathbb{R})$ is VP_o iff $A \geq 0$ or $A \leq 0$ or there exist nonzero vectors u, $w \in \mathbb{R}^n$ with $u \geq 0$ such that $Ax = (w^T x)u$ for all $x \in \mathbb{R}^n$.

Proof. The "if" part is obvious. Therefore, let A be VP_o but neither nonnegative nor nonpositive. Then we claim that there exist x, $y \in \mathbb{R}^n$ both nonnegative such that $Ax \geq 0$ while $Ay \leq 0$ with neither Ax nor Ay being a null vector. By our assumptions, there exist p, q, r, s such that $a_{pq} > 0$ and $a_{rs} < 0$. Since A is VP_o and $a_{pq} > 0$ it follows that $Ae_q \geq 0$. Similarly $Ae_s \leq 0$. Moreover $(Ae_q)_p$ and $(Ae_s)_r$ are nonzero. Hence our claim is true.

Now let x,y be any such two vectors satisfying our claim. For $c \in (0,\infty)$, our choice of x,y shows that $cx+y \geq 0$. Hence cAx+Ay is nonnegative for sufficiently large values of c and nonpositive for sufficiently small values of c. Define $c_o = \sup\{c: cAx+Ay \leq 0\}$. Clearly c_o exists and is positive. Moreover for $\varepsilon>0$,

$$(c_o-\varepsilon)Ax + Ay \leq 0 \tag{3.3.4}$$

and

$$(c_o+\varepsilon)Ax + Ay \geq 0. \tag{3.3.5}$$

Together (3.3.4) and (3.3.5) imply that

$$-\varepsilon Ax \leq c_oAx+Ay \leq \varepsilon Ax. \tag{3.3.6}$$

Since $\varepsilon>0$ is arbitrary, it follows that

$$c_oAx + Ay = 0. \tag{3.3.7}$$

Notice that c_o depends on x and y. Replacing c_o by $c(x,y)$ and fixing x as e_q in (3.3.7) we find that

$$Ay = -c(e_q,y)Ae_q \tag{3.3.8}$$

and hence

$$Ay = \alpha(y)u \tag{3.3.9}$$

with $u = Ae_q$ and $\alpha(y)$ being a scalar depending on y. If we take y as e_s in (3.3.7) then

$$\begin{aligned} Ax &= -\frac{1}{c(x,e_s)}\,Ae_s \\ &= \frac{c(e_q,e_s)}{c(x,e_s)}\,Ae_q. \end{aligned}$$

Hence,

$$Ax = \alpha(x)u. \tag{3.3.10}$$

By VP_0 property of A, any nonnegative $z \in \mathbb{R}^n$ is of the type x (i.e. with $Az \geq 0$ and $Az \neq 0$) or of the type y (i.e. with $Az \leq 0$ and $Az \neq 0$) or such that $Az = 0$. Hence it is clear that for any $z \geq 0$ in $\mathbb{R}^n$, Az is of the form $\alpha(z)u$ for some fixed $u \geq 0$ in $\mathbb{R}^n$ with $u \neq 0$. Since any vector can be expressed as the difference of two nonnegative vectors, the very same form given in (3.3.10) holds for any $x \in \mathbb{R}^n$. Since $u_p = (Ae_q)_p = a_{pq} \neq 0$, by equating the p-th coordinates of Ax and $\alpha(x)u$, we have

$$A_{(p)}x = \alpha(x)a_{pq}$$

and hence

$$\alpha(x) = (a_{pq})^{-1}A_{(p)}x. \tag{3.3.11}$$

Now by setting $w^T = (a_{pq})^{-1}A_{(p)}$, the theorem follows. ▲▲▲

As an almost immediate consequence of this theorem we have the following well-known corollary.

COROLLARY 3.3.1. Let $A \in M_n(\mathbb{R})$. Then $Ax \geq 0$ for all $x \geq 0$ in $\mathbb{R}^n$ iff $A \geq 0$.

In the sequel while discussing VP_2 matrices in $M_n(\mathbb{R})$ we assume that $n \geq 3$. Similarly for VP_1 matrices we take $n \geq 2$. The reason for these restrictions is obvious.

In order to prove our main theorem of this section, we require the following two lemmas.

LEMMA 3.3.1. Let B be row stochastic and VP_2. Then B is consistently monotone-preserving (CMP).

Proof. First we shall prove that B is monotone-preserving. Without loss of generality, it is sufficient to consider an arbitrary nondecreasing vector $x = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$, and show that Bx is monotonic. If $x_1 = x_n$, then x is a constant multiple of e and in this case Bx is trivially monotonic. So we assume that $x_1 < x_n$. Now

$$v(x-ce) = \begin{cases} 1 \text{ if } x_1 < c < x_n, \\ 0 \text{ otherwise.} \end{cases}$$

Since Be=e and B is VP_2, it follows that

$$v(Bx-ce) = \begin{cases} 1 \text{ if } x_1 < c < x_n, \\ 0 \text{ otherwise.} \end{cases}$$

If Bx is not monotonic, it will have at least one set of three coordinates $(Bx)_k$, k=p,q,r (p<q<r) such that

$$(Bx)_q > M \text{ or } (Bx)_q < m$$

where

$$M = \max\{(Bx)_p, (Bx)_r\}$$

$$m = \min\{(Bx)_p, (Bx)_r\}.$$

Now choosing c with $M < c < (Bx)_q$ or $(Bx)_q < c < m$ depending on the situation, we obtain a contradiction that $v(Bx-ce) \geq 2$. Hence Bx is monotonic.

Next, we shall establish that B is CMP. If it is not so, then there will exist nondecreasing vectors $x = (x_1,\ldots,x_n)^T$ and $y = (y_1,\ldots,y_n)^T$ with $x_1 < x_n$ and $y_1 < y_n$ such that

$$(Bx)_1 < (Bx)_n \text{ and } (By)_1 > (By)_n.$$

Clearly for any $c \in (0,\infty)$, $cx+y$ is $\uparrow$. This implies that $cBx+By$ is monotonic. Moreover it is nondecreasing for sufficiently large values of c and nonincreasing for sufficiently small values of c. Let $c_o = \sup\{c: cBx+By \text{ is } \downarrow\}$ Then c_o exists and is positive. Evidently, for $\varepsilon>0$,

$$(c_o-\varepsilon)Bx + By \text{ is } \downarrow \text{ and } (c_o+\varepsilon)Bx + By \text{ is } \uparrow,$$

so that for $k=1,\ldots,n-1$

$$((c_o-\varepsilon)Bx + By)_k \geq ((c_o-\varepsilon)Bx + By)_{k+1} \tag{3.3.12}$$

and

$$((c_o+\varepsilon)Bx + By)_k \leq ((c_o+\varepsilon)Bx + By)_{k+1}. \tag{3.3.13}$$

From (3.3.12) and (3.3.13) we get

$$-\varepsilon\{(Bx)_{k+1}-(Bx)_k\} \leq (c_oBx+By)_k-(c_oBx+By)_{k+1}$$

$$\leq \varepsilon\{(Bx)_{k+1}-(Bx)_k\}. \tag{3.3.14}$$

Since ε is arbitrary it follows at once that

$$c_oBx + By = \alpha e \tag{3.3.15}$$

where α is a real constant depending on x and y. Indeed, c_o also depends on x and y. Let us replace c_o by $c(x,y)$ and α by $\alpha(x,y)$ in (3.3.15) so that

$$By = \alpha(x,y)e - c(x,y)Bx. \tag{3.3.16}$$

In this relation, fixing x as x_o (which could be any vector with $(x_o)_1 < (x_o)_n$ and $(Bx_o)_1 < (Bx_o)_n$) we have

$$By = \alpha(y)e - \beta(y)u \tag{3.3.17}$$

where $\alpha(y) = \alpha(x_o,y)$, $\beta(y) = c(x_o,y)$ and $u = Bx_o$. Similarly, fixing y as y_o (which could be any vector with $(y_o)_1 < (y_o)_n$ and $(By_o)_1 > (By_o)_n$) in (3.3.16) and substituting for By_o

obtained from (3.3.17) we find after simplifications that there exist scalars $\alpha(x)$ and $\beta(x)$ such that

$$Bx = \alpha(x)e - \beta(x)u. \tag{3.3.18}$$

If $x_1 < x_n$ with $(Bx)_1 = (Bx)_n$ or if $x_1=x_n$ trivially Bx is of the above form since Bx is a constant multiple of e. From the above arguments it is clear that if x is $\uparrow$, then in any case Bx assumes the form given in (3.3.18). Furthermore, it is well known that any real vector can be expressed as a difference of two nondecreasing vectors. Hence for any $x \in \mathbb{R}^n$, Bx assumes the same form given in (3.3.18) which is always monotonic since u is monotonic. This implies that for any $x \in \mathbb{R}^n$, $v(Bx) \leq 1$ which is contrary to the VP_2 property of B, since (as $n \geq 3$) there are vectors in $\mathbb{R}^n$ with variation 2. This contradiction establishes the CMP property of B.▲▲▲

LEMMA 3.3.2. Let B be row stochastic, VP_2 and trend-preserving. Then B=I.

Proof. We shall first show that each column of $B = (b_{ij})$ has at least one entry as 1. In this process, we also exhibit that $b_{11} = b_{nn} = 1$. Define $f_j = e-e_j$ for j=1,...,n. As Be=e, for a real c, obviously

$$v(Bf_1-ce) = v(f_1-ce) = \begin{cases} 1 \text{ if } 0 < c < 1, \\ 0 \text{ otherwise.} \end{cases} \tag{3.3.19}$$

Since e_1 is $\downarrow$, clearly $B^{(1)}$ is $\downarrow$. Moreover, $B^{(1)} \geq 0$. If $b_{11} \neq 1$, let it be α so that $0 \leq \alpha < 1$. The case $\alpha=0$ leads to the situation $B^{(1)} = 0$ and we have $1 = v(e_1-e_2) = v(Be_1-Be_2) = v(-B^{(2)}) = 0$, a contradiction. Hence $0 < \alpha < 1$.

Since (f_1-ce) is $\uparrow$, (Bf_1-ce) is $\uparrow$. Now by choosing c such that $0 < c < 1-\alpha$, we see that $(Bf_1-ce)_1 = 1-c-\alpha > 0$. consequently $v(Bf_1-ce) = 0$ which violates (3.3.19). Hence $b_{11} = 1$. By a similar analysis it can be easily shown that $b_{nn} = 1$.

Our next claim is that for each $j=2,\ldots,n-1$, $b_{ij} = 1$ for some $i=i(j)$. Let $j \in \{2,\ldots,n-1\}$. It may be observed that

$$v(Bf_j-ce) = v(f_j-ce) = \begin{cases} 2 \text{ if } 0 < c < 1, \\ 0 \text{ otherwise.} \end{cases} \qquad (3.3.20)$$

If $b_{ij} \neq 1$ for all $i=1,\ldots,n$, then $0 < \beta \leq 1$, where $\beta = \min_i(1-b_{ij})$. Now for any c satisfying $0 < c < \beta$, we have $Bf_j-ce > 0$ and so $v(Bf_j-ce) = 0$ which contradicts (3.3.20) and hence for each $j \in \{2,\ldots,n-1\}$, $b_{ij} = 1$ for some $i=i(j)$.

We have established that each column of B has at least one entry as 1. In view of $B \geq 0$ and $\sum_{i,j} b_{ij} = n$, it easily follows that each column of B has precisely one nonzero entry 1. Since every row sum of B is 1, it further follows that B is a permutation matrix. Finally, by considering $x = (1,\ldots,n)^T$, we use the trend-preserving property of B to conclude that B=I. ▲▲▲

We are now prepared to establish the following main result of this section which gives a complete characterization of VP_2 matrix transformations.

THEOREM 3.3.4. Let $A \in M_n(\mathbb{R})$ with $n \geq 3$. Then the following three statements are equivalent.

(i) A is VP_2.

(ii) A is VP.

(iii) A is a diagonal (or skew-diagonal) matrix having only positive or only negative entries in the diagonal (or skew-diagonal). That is to say, A has one of the four forms

$$\pm \begin{bmatrix} p_1 & & \\ & p_2 & \\ & & \ddots \\ & & & p_n \end{bmatrix}, \quad \pm \begin{bmatrix} & & & p_n \\ & & \iddots & \\ & p_2 & & \\ p_1 & & & \end{bmatrix}$$

where $p_i > 0$, $i=1,\ldots,n$.

Proof. (iii) ==> (ii) ==> (i) is trivial. We shall therefore prove that (i) ==> (iii). Assuming (i), we have in particular that A is VP_0. Hence by Theorem 3.3.3, either $A \geq 0$, or $A \leq 0$, or there exist nonzero vectors $u,w \in \mathbb{R}^n$ with $u \geq 0$ such that $Ax = (w^T x)u$ for all $x \in \mathbb{R}^n$. For the last mentioned possibility and as well as for the case $A=0$, we have $v(Ax) = 0$ for all $x \in \mathbb{R}^n$ whereas there exists $x \in \mathbb{R}^n$ $(n \geq 3)$ with $v(x) = 1$ or 2. This contradicts the VP_2 property of A. Hence $A \geq 0$, or $A \leq 0$ and has at least one nonzero entry. Denoting the i-th row sum of A by s_i, $i=1,\ldots,n$, let us construct a new matrix B as follows.

Let $x \in \mathbb{R}^n$ be arbitrary. For k=1,...,n define

$$(Bx)_k = \begin{cases} (Ax)_k/s_k & \text{if } s_k \neq 0, \\ (Bx)_{k-p} & \text{if } s_k = 0. \end{cases}$$

where p is an integer with smallest possible $|p|$ satisfying $s_{k-p} \neq 0$. If two such p's are available, take the positive one. Thus, whenever $s_k \neq 0$, the k-th row of B is

$$\left(\frac{a_{k1}}{s_k}, \frac{a_{k2}}{s_k}, \ldots, \frac{a_{kn}}{s_k} \right)$$

and after defining all such rows, a row of B corresponding to a vanishing s_k is taken as the one which is nearest to its position among the rows already defined for nonvanishing row sums. If there are two such nearest rows any one can be chosen. However for the sake of definiteness, choose the preceding one among the two.

An immediate observation about B is that $B \geq 0$ and is row stochastic. Another interesting fact is that B is VP_2. To see this, first let us consider the case $A \geq 0$. Whenever $s_k \neq 0$, for any $x \in \mathbb{R}^n$,

$$\operatorname{sgn}((Ax)_k) = \operatorname{sgn}((Bx)_k)$$

where sgn(α) (read: signum α), for a real α, is defined as

$$\operatorname{sgn}(\alpha) = \begin{cases} 1 & \text{if } \alpha > 0, \\ 0 & \text{if } \alpha = 0, \\ -1 & \text{if } \alpha < 0. \end{cases}$$

If $s_k = 0$, obviously $(Ax)_k = 0$. On the other hand $(Bx)_k$ may admit any signum but it will be in such a manner that $v(Bx) = v(Ax)$. This will be clear from the following arguments:

(1) If $s_k = 0$ for $k=1,\ldots,r$ $(r<n)$ and $s_{r+1} \neq 0$ then all $(Bx)_k$, $k=1,\ldots,r$ admit the same signum as that of $(Bx)_{r+1}$ and hence,

$$v\{(Bx)_1,\ldots,(Bx)_r,(Bx)_{r+1}\} = 0.$$

(2) If $s_k = 0$ for $k=m+1,\ldots,n$ $(m>0)$ and $s_m \neq 0$ then all $(Bx)_k$, $k=m+1,\ldots,n$ admit the same signum as that of $(Bx)_m$. Hence,

$$v\{(Bx)_m,(Bx)_{m+1},\ldots,(Bx)_n\} = 0.$$

(3) If $s_k = 0$ for $k=m+1,\ldots,r$ where $m > 0$ and $r < n$ with $s_m \neq 0$ and $s_{r+1} \neq 0$, then the first $[(r-m+1)/2]$ terms in the sequence $(Bx)_{m+1}, \ldots, (Bx)_r$ will have the same signum as that of $(Bx)_m$ and the remaining terms will have the signum as that of $(Bx)_{r+1}$ where $[(r-m+1)/2]$ denotes the integral part of $(r-m+1)/2$. Hence,

$$v\{(Bx)_m,(Bx)_{m+1},\ldots,(Bx)_r,(Bx)_{r+1}\} = v\{(Bx)_m,(Bx)_{r+1}\}.$$

Thus, since the possibilities (1)-(3) are exhaustive, in any case the inclusion of coordinates of Bx corresponding to vanishing row sums of A does not affect the variation in the sequence of coordinates of Bx corresponding to nonvanishing row sums. Hence we have $v(Bx) = v(Ax)$. Since B is the same for both A and -A, we have for $A \leq 0$, $v(Bx) = v(-Ax) = v(Ax)$. Since A is VP_2, it now follows that B is also VP_2.

Now by Lemma 3.3.1, B is CMP. Without loss of generality we can assume that B is trend-preserving. Otherwise, i.e., if B is trend-reversing, we consider $\tilde{D}B$ in

the validity of $VP_2 \Longleftrightarrow VP$ in the discrete case as in the continuous case, now it is quite natural to seek for some counterexamples as above in the discrete case also. Therefore, we tried to construct some matrices which are VP_1 but not VP_2 and we could not succeed in our attempt. On the contrary, to our surprise, we were subsequently able to establish that also VP_1 implies VP in the discrete case. The proof of this interesting result depends on the following basic lemma which is a refinement of Lemma 3.3.1 with some weaker hypothesis.

LEMMA 3.4.1. Let B be row stochastic and VP_1. Then B is CMP.

Proof. The monotone-preserving property of B follows as in the proof of Lemma 3.3.1. Since we are not allowed to use VP_2 property, there is no use in continuing further as in Lemma 3.3.1. Therefore, let us argue in a different manner.

Consider $g = (0,1,\ldots,n-1)^T$. Then

$$v(Bg-ce) = v(g-ce) = \begin{cases} 1 \text{ for } 0 < c < n-1, \\ 0 \text{ otherwise.} \end{cases} \tag{3.4.1}$$

From this it is clear that Bg cannot be a multiple of e. Let $m = \min_i (Bg)_i$ and $M = \max_i (Bg)_i$. We shall prove that $m=0$ and $M=n-1$. Since $(Bg)_i \geq 0$ for $i=1,\ldots,n$, obviously $m \geq 0$. If $m \neq 0$, then for sufficiently small c satisfying $0 < c < m$, we have $v(Bg-ce) = 0$, a contradiction. Hence $m=0$.

If $M > n-1$, then for $c=n-1$, we see that $v(Bg-ce) = 1$, contrary to (3.4.1). Again if $M < n-1$, for c with

$M < c < n-1$, we have a contradiction, namely $v(Bg-ce) = 0$. These arguments assert that $M=n-1$.

Now, in view of the monotonicity of Bg, it follows immediately that either

$$(Bg)_1 = 0 \text{ and } (Bg)_n = n-1 \tag{3.4.2}$$

or

$$(Bg)_1 = n-1 \text{ and } (Bg)_n = 0. \tag{3.4.3}$$

For any $x = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$, it can be easily shown that

$$-2\|x\|_\infty g \leq x-x_1 e \leq 2\|x\|_\infty g \tag{3.4.4}$$

and

$$-2\|x\|_\infty((n-1)e-g) \leq x-x_n e \leq 2\|x\|_\infty((n-1)e-g) \tag{3.4.5}$$

where

$$\|x\|_\infty = \max_i |x_i| .$$

Since $B \geq 0$ and $Be=e$, it follows from (3.4.4) that

$$-2\|x\|_\infty Bg \leq Bx-x_1 e \leq 2\|x\|_\infty Bg. \tag{3.4.6}$$

Therefore, if we assume (3.4.2), then by considering the first coordinates in the vector relation (3.4.6), we have $(Bx)_1 = x_1$. Similarly by considering the n-th coordinates in the vector relation obtained by pre-multiplying (3.4.5) by B, we get $(Bx)_n = x_n$.

On the other hand if we assume (3.4.3), then in a similar manner it easily follows that $(Bx)_n = x_1$ and $(Bx)_1 = x_n$. The above arguments hold in particular for any nondecreasing vector x. Hence we have shown that x is $\uparrow$ always implies Bx is $\uparrow$ or always implies Bx is $\downarrow$. This completes the proof of the lemma. ▲▲▲

We can now state and prove the main theorem of this section.

THEOREM 3.4.1. $A \in M_n(\mathbb{R})$ $(n \geq 2)$ is VP_1 iff A is a diagonal (or skew-diagonal) matrix having all positive or all negative diagonal (or skew-diagonal) entries.

Proof. The sufficiency part is immediate. To establish the necessity part we use the method of induction. First of all, we shall get rid of the case n=2.

Since, in particular, A is VP_0, by the same arguments given in the earlier part of the proof of (i) ==> (iii) in Theorem 3.3.4, A is either nonnegative or nonpositive with at least one nonzero entry. Hence A is always of the form $\pm\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with a, b, c, d ≥ 0. Consider $x = \begin{bmatrix} p \\ -1 \end{bmatrix}$ with $p > 0$ so that $Ax = \pm \begin{bmatrix} ap-b \\ cp-d \end{bmatrix}$. Since $v(x)=1$, it follows that $v(Ax)=1$ implying, in either case, that

$$(ap-b)(cp-d) < 0. \tag{3.4.7}$$

If $ac \neq 0$, then for sufficiently large values of p, (3.4.7) does not hold. Hence ac=0. Similarly if $bd \neq 0$, then for sufficiently small values of p, (3.4.7) does not hold. Hence bd=0. If a=0 then either c or b being zero will lead to a contradiction. Hence if a=0, then $c \neq 0$ and $b \neq 0$ which in turn implies that d=0. On the other hand, if $a \neq 0$, then c=0 which implies that $d \neq 0$ and hence b=0. This establishes the necessity part for the case n=2.

Before applying the induction process, once again we proceed along the same lines as in the proof of (i) ==> (iii) in Theorem 3.3.4 to construct the matrix B as explained over

there. By the arguments in that proof, B is row stochastic and VP_1 since A is VP_1 and $v(Bx) = v(Ax)$ for all $x \in \mathbb{R}^n$. Now by Lemma 3.4.1, we know that B is CMP. As in Theorem 3.3.4, there is no loss of generality in assuming that B is TRP, since otherwise we can consider $\tilde{D}B$ in the place of B where $\tilde{D}$ is the skew-diagonal matrix defined by (3.3.21). In Lemma 3.3.2, in fact we have used only VP_1 property of B to prove that $b_{11} = b_{nn} = 1$ and $B = (b_{ij})$ is assumed to be row stochastic and TRP. Hence B can be expressed in the partitioned form $\begin{bmatrix} 1 & 0 \\ v & E \end{bmatrix}$ where $E \in M_{n-1}(\mathbb{R})$ and $v \in \mathbb{R}^{n-1}$. For any $y \in \mathbb{R}^{n-1}$,

$$B\begin{bmatrix} 0 \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ v & E \end{bmatrix} \begin{bmatrix} 0 \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ Ey \end{bmatrix}.$$

It shows clearly that E is VP_1 on $\mathbb{R}^{n-1}$. Now to begin the induction process let us assume that the necessary part of the theorem under consideration is true for (n-1)-dimensional case. Since the bottom right corner element of E is already 1, E cannot be skew-diagonal and hence we must have

$$B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \alpha_2 & q_2 & 0 & \cdots & 0 & 0 \\ \alpha_3 & 0 & q_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \alpha_{n-1} & 0 & 0 & \cdots & q_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

where $\alpha_i \geq 0$, $q_i > 0$, $i=2,\ldots,n-1$. Now consider $x = ce_1 - e_2$ with $c > 0$. Clearly $v(x)=1$. By direct computations,

$$Bx = \begin{bmatrix} c \\ \alpha_2 c - q_2 \\ \alpha_3 c \\ \vdots \\ \alpha_{n-1} c \\ 0 \end{bmatrix}.$$

Since $c > 0$, and $\alpha_i c \geq 0$ for $i=3,\ldots,n-1$, a necessary condition for $v(Bx) = 1$ is

$$\alpha_2 c - q_2 < 0. \tag{3.4.8}$$

Moreover if any one of $\alpha_3,\ldots,\alpha_{n-1}$ is nonzero, by virtue of (3.4.8), $v(Bx) > 1$. Hence $\alpha_3 = \ldots = \alpha_{n-1} = 0$. If $\alpha_2 \neq 0$ then c could be chosen large enough so that (3.4.8) is not true. Hence $\alpha_2=0$. Since $Be=e$, it follows that $B=I$. Hence by the same arguments as in Theorem 3.3.4, we conclude that A is having any one of the four forms given in the statement of the present theorem. We have already shown that the result is true in the 2x2 case. This completes the proof of the theorem. ▲▲▲

REMARK 3.4.1. The theorem just proved gives the structure of VP_1 matrix transformations and from this we infer that VP_m <==> VP for $m=1,\ldots,n-1$. In particular, Theorem 3.3.4 follows from Theorem 3.4.1. It may also be noted that any nonnegative or nonpositive matrix which is not of the four forms given in Theorem 3.4.1 is VP_0 but fails to be VP_1. Thus, in the case of matrix transformations

1 is the smallest integer m such that $VP_m \Longrightarrow VP$. This is quite startling in the light of the corresponding result in the continuous case in which m=2 is the smallest such integer. Also a comparison of the proofs of Theorems 3.3.4 and 3.4.1 clearly indicates the simplicity and the power of induction arguments, whenever applicable.

We conclude this section by giving a characterization of trend-preserving matrices. We have already used the concept of trend-preserving matrices in the study of variation-preserving matrices. We recall that $A \in M_n(\mathbb{R})$ is trend-preserving if $(Ax)_1 \leq \ldots \leq (Ax)_n$ for all $x = (x_1,\ldots,x_n)^T \in \mathbb{R}^n$ satisfying $x_1 \leq \ldots \leq x_n$.

A special interest in trend-preserving matrices arises due to the interpretation of stochastic matrices as diffusion processes where it is of natural interest to know whether the next stage keeps a trend the same.

THEOREM 3.4.2. $A = (a_{ij}) \in M_n(\mathbb{R})$ is trend-preserving iff

$$\sum_{j=1}^{k} (a_{i+1,j} - a_{ij}) \leq 0, \quad i,k=1,\ldots,n-1 \qquad (3.4.9)$$

and every row of A has the same sum.

Proof. Let $x = (x_1,\ldots,x_n)^T$ be such that $x_1 \leq \ldots \leq x_n$. If A is trend-preserving, then for i=1,...,n-1

$$(Ax)_i \leq (Ax)_{i+1}.$$

Hence,

$$\sum_{j=1}^{n} (a_{ij} - a_{i+1,j})x_j \leq 0, \quad i=1,\ldots,n-1. \qquad (3.4.10)$$

Now setting $x = -(e_1+e_2+\ldots+e_k)$ in (3.4.10) for $k=1,\ldots,n-1$, we have (3.4.9). Next, we shall put $x = -e$ in (3.4.10) so that

$$\sum_{j=1}^{n} (a_{i+1,j} - a_{ij}) \leq 0. \tag{3.4.11}$$

Similarly if we put $x = e$ in (3.4.10), then

$$\sum_{j=1}^{n} (a_{i+1,j} - a_{ij}) \geq 0. \tag{3.4.12}$$

From (3.4.11) and (3.4.12), it follows that

$$\sum_{j=1}^{n} a_{i+1,j} = \sum_{j=1}^{n} a_{ij}, \tag{3.4.13}$$

i.e., every row of A has the same sum.

Next, we shall assume that (3.4.9) and (3.4.13) are true. Let $x_1 \leq \ldots \leq x_n$. Simple calculations show that

$$\begin{aligned}(Ax)_{i+1} - (Ax)_i &= \sum_{j=1}^{n}(a_{i+1,j} - a_{ij})x_j \\ &= \sum_{j=1}^{n}\{\sum_{k=1}^{j}(a_{i+1,k}-a_{ik}) - \sum_{k=1}^{j-1}(a_{i+1,k}-a_{ik})\}x_j \\ &= \sum_{j=1}^{n-1}\{(x_j-x_{j+1})\sum_{k=1}^{j}(a_{i+1,k}-a_{ik})\} \\ &\qquad + x_n \sum_{k=1}^{n}(a_{i+1,k}-a_{ik})\end{aligned}$$

which turns out to be nonnegative in view of our assumptions. It may be observed that the last term vanishes because of (3.4.13). The proof is now complete. ▲▲▲

It may be noted that the conditions (3.4.9) are equivalent to saying that each one of the (n-1) vectors $\sum_{j=1}^{k} A^{(j)}$, $k=1,\ldots,n-1$, $A^{(j)}$ denoting the j-th column of A, is nonincreasing.

An obvious consequence of the above theorem is the following.

COROLLARY 3.4.1. $A \in M_n(\mathbb{R})$ is trend-reversing iff

$$\sum_{j=1}^{k} (a_{i+1,j} - a_{ij}) \geq 0, \quad i,k=1,\ldots,n-1 \qquad (3.4.14)$$

and every row of A has the same sum.

3.5. Inertia-, Angularity- and Zero-preserving Transformations on $\mathbb{C}^n$.

In the present section we confine our attention to matrix transformations on $\mathbb{C}^n$. We shall determine the structure of complex matrices preserving inertia, angularity and the number of zero coordinates of vectors in $\mathbb{C}^n$. Amongst other problems of a similar nature involving transformations mapping $\mathbb{C}^n$ into intself which have already been solved, we may recall the well-known result that $A: \mathbb{C}^n \to \mathbb{C}^n$ is norm-preserving iff it is unitary [18, p.230].

We require, to begin with, the following definitions.

DEFINITION 3.5.1 (see Bahl and Cain [26]). The inertia $In(x)$ of a vector $x \in \mathbb{C}^n$ is defined as the ordered triple $(\pi(x), \nu(x), \delta(x))$ where the entries in the triple denote respectively the number of coordinates in x with positive, negative, and zero real parts. In other words, if $x = (x_1,\ldots,x_n)^T$, then $In(x) = In(D)$ where $D = diag(x_1,\ldots,x_n)$. Similarly the angularity $\theta[x]$ of x can be identified with the angularity of $D = diag(x_1,\ldots,x_n)$, i.e., $\theta[x] = \theta[D]$.

In the sequel, e_i, e and $A^{(j)}$ will have the same meaning as used in the previous sections. It may be noted that $In(e) = (n,0,0)$ and $In(e_j) = (1,0,n-1)$ for $j=1,\ldots,n$.

DEFINITION 3.5.2. $A \in M_n$ is said to be inertia-preserving if $In(Ax) = In(x)$ for all $x \in \mathbb{C}^n$; it is said to be angularity-preserving if $\theta[Ax] = \theta[x]$ for all $x \in \mathbb{C}^n$.

It may be observed that A preserves inertia whenever it preserves angularity.

DEFINITION 3.5.3. $A \in M_n$ is said to be zero-preserving if $Z(Ax) = Z(x)$ for every $x \in \mathbb{C}^n$ where $Z(x)$ denotes the number of zero coordinates in the vector x.

The definitions of permutation matrix, generalized permutation matrix (g.p.m.), and nonnegative g.p.m. are already given in Section 1.3 and these will be required in the following results.

The first result of the section characterizes inertia-preserving matrices. It will also be of use in proving some results in the next section.

THEOREM 3.5.1. $A \in M_n$ is inertia-preserving iff A is a nonnegative generalized permutation matrix.

Proof. The "if" part is obvious. On the other hand, if $In(Ax) = In(x)$ for every $x \in \mathbb{C}^n$, then by considering $x = e_j$ we immediately conclude that each column of A has precisely one entry with positive real part and the remaining

(n-1) entries are having zero real parts. Similarly by considering $x = \sqrt{-1}\, e_j$ we see that each entry of $\sqrt{-1}\, A^{(j)}$, $j=1,\ldots,n$ has zero real part implying that A is real.

Combining the above two observations, it follows that each column of A has precisely one positive entry, the other entries being zero. If any row of A has more than one positive entry then at least one row of A will have all zero entries so that $(n,0,0) = In(e) = In(Ae) \neq (n,0,0)$. This contradiction completes the proof of the theorem. ▲▲▲

As an immediate consequence of the above theorem, we have

COROLLARY 3.5.1. $A \in M_n$ is angularity-preserving iff A is inertia-preserving.

Also we have the following corollary.

COROLLARY 3.5.2. Let $A \in M_n$ be such that for each $x \in \mathbb{C}^n$, there is a permutation matrix P_x such that $Ax = P_x x$. Then there is a fixed permutation matrix P such that $Ax = Px$ for every $x \in \mathbb{C}^n$.

Proof. From the hypothesis, it is clear that A preserves the coordinates of x but in a permuted form. Therefore, clearly A is inertia-preserving and hence it is a nonnegative g.p.m.. Since each of the entries of Ae is 1, it follows that A is a permutation matrix. ▲▲▲

THEOREM 3.5.2. $A \in M_n$ is zero-preserving iff A is a generalized permutation matrix.

Proof. The sufficiency part of the theorem is obvious. To prove the other part, we see that $Z(A^{(j)}) = Z(Ae_j) = Z(e_j) = n-1$, showing that each column of A has precisely one nonzero entry. If any row of A has more than one nonzero entry, then there is at least one row of A having all zero entries. In this case, we have the contradiction $0 = Z(e) = Z(Ae) \neq 0$. The proof is now complete. ▲▲▲

3.6. Inertia- and Angularity-preserving Transformations on Circulants.

A matrix of the form

$$C = \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & \cdots & c_{n-2} \\ \vdots & \vdots & & \vdots \\ c_1 & c_2 & \cdots & c_0 \end{bmatrix} \qquad (3.6.1)$$

in which each row is obtained from the preceding row by shifting the elements cyclically one column to the right is called a circulant matrix (see Lancaster [13, p.267], Mirsky [18, p.432]) or simply a circulant (see Bellman [4, p.242], Marcus and Minc [16, p.66]). A circulant matrix is also known as a cyclic matrix [285]. Since the term "circulant" was originally used to refer to the determinant of a matrix of the form given in (3.6.1), Good [291] called a circulant matrix by the name circulix. Several authors (e.g. Varga [306]), define a circulant matrix as a matrix

of the form

$$\begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_1 & c_2 & \cdots & c_0 \\ \vdots & \vdots & & \vdots \\ c_{n-1} & c_0 & \cdots & c_{n-2} \end{bmatrix} .$$

For some more references, regarding the two definitions of circulant matrices, one may refer to the survey paper by Roebuck and Barnett [302].

Throughout our study in the thesis, by a circulant we mean only a matrix of the form given in (3.6.1) and we follow Davis [287,288] in denoting this matrix by $\mathrm{circ}(c_0,c_1,\ldots,c_{n-1})$. We denote the class of all $n \times n$ circulants by C_n. In the sequel D_n is used to denote the class of all $n \times n$ diagonal matrices. F denotes the discrete Fourier transform (DFT) matrix defined by $F = (n^{-1/2}\omega^{-(j-1)(k-1)})$, $j,k=1,\ldots,n$, where $\omega = \exp(2\pi i/n)$ with $i = \sqrt{-1}$. It may be noted that F is unitary and symmetric. In the present section P is reserved to denote the special circulant $\mathrm{circ}(0,1,0,\ldots,0)$.

The interest in the study of circulants stems from their applications in various fields. They arise in the numerical solution of boundary value problems by boundary contraction [298] and are also used to approximate and to explain the behaviour of Toeplitz matrices which have

numerous applications in information theory and estimation theory [293]. Dynamical systems involving circulants arise in the study of molecular self-organization and also in the evolution of animal behaviour [303]. For some more applications, especially in solid-state physics and statistics, one may refer to the recent paper by Searle [304]. Moreover, the advent of fast Fourier transform (FFT) methods [5] has greatly enhanced the use of circulant approximations [295] in various fields.

Circulants are interesting geometrically and simple to handle algebraically [287]. They have many desirable and nice properties as listed in Searle [304]. In particular, we note that under the usual matrix addition and multiplication, C_n is a commutative ring isomorphic to the ring D_n of diagonal matrices [284]. In fact,

$$\begin{aligned} \operatorname{circ}(c_0, c_1, \ldots, c_{n-1}) &= F^* \operatorname{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1}) F \\ &= F \operatorname{diag}(\lambda_0, \lambda_{n-1}, \ldots, \lambda_1) F^* \end{aligned}$$

where

$$\lambda_k = \sum_{j=0}^{n-1} c_j \exp(2j\pi ki/n), \quad k=0,1,\ldots,n-1.$$

Since F is unitary, it follows that $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ are the eigenvalues of $\operatorname{circ}(c_0, c_1, \ldots, c_{n-1})$, that circulants are normal and further that whenever nonsingular they have a circulant inverse. It may also be observed that the eigenvalues of P are $1, \omega, \omega^2, \ldots, \omega^{n-1}$.

In this section, we first present a basic theorem incorporating various characterizations of circulants, collected from different sources. Next, we determine the set of all nonsingular matrices S such that S^*CS is a circulant for every circulant C. Finally, based on some of the results proved in the preceding section, we characterize the structure of the linear transformations $T: M_n \to M_n$ such that

(a) $T(C_n) \subseteq C_n$ and $In(T(C)) = In(C)$ for all $C \in C_n$ and

(b) $T(C_n) \subseteq C_n$ and $\Theta[T(C)] = \Theta[C]$ for all $C \in C_n$.

The characterizations of circulants are summarized as

THEOREM 3.6.1. Let $C \in M_n$ and P and F denote respectively circ(0,1,0,...,0) and DFT matrix. Then the following conditions are all equivalent.

(i) $C \in C_n$.

(ii) C is a polynomial of degree $\leq n-1$, in P.

(iii) C is a polynomial in P.

(iv) $ACB \in C_n$ for any nonsingular A and B in C_n.

(v) C commutes with P.

(vi) $FCF^* \in D_n$

(vii) $F^*CF \in D_n$

(viii) The columns of F^* are eigenvectors of C.

(ix) The columns of F are eigenvectors of C.

(x) $F^2C(F^*)^2 \in C_n$.

(xi) $(F^*)^2CF^2 \in C_n$.

Proof. It is well known that $\mathrm{circ}(c_0, c_1, \ldots, c_{n-1}) = \sum_{j=0}^{n-1} c_j P^j$, $P^n = I$ and $P^r = P^s$ for $r \equiv s \pmod{n}$. Thus, (i) ==> (ii) ==> (iii) ==> (i). The equivalence of (i) and (iv) follows from the closure property of C_n under usual multiplication. The proof of (i) <==> (v) is given in Ablow and Brenner [279] and also in Brenner [283]. The equivalence of (i), (vi) and (viii) is part of the substance of Chalkley's paper [284].

A quick way of seeing that (i) <==> (vii) is to consider C^T which is a circulant if C is so [308]. Now (i) <==> $FCF^* \in D_n$ <==> (vii) since $F^T = F$. Clearly (vii) <==> (ix). It is easy to see that (x) <==> (i) <==> (xi). In fact, if $C \in C_n$ then $F^2C(F^*)^2 = (F^*)^2CF^2 = C^T$. ▲▲▲

In Theorem 2.4.5, we characterized the class of all nonsingular matrices C such that C^*AC is normal for every normal A. Now we shall consider a similar problem for circulants. For this purpose, we shall first prove three simple lemmas.

LEMMA 3.6.1. If $S \in M_n$ is nonsingular and $S^{-1}PS \in C_n$ then $S^*S \in C_n$.

Proof. Since $S^{-1}PS \in C_n$, $F(S^{-1}PS)F^* = D \in D_n$. This shows that the columns of SF^* are eigenvectors of P. Since the eigenvalues of P are all distinct and the eigenvectors of P are columns of F also, it follows that $SF^* = F^*Q$ for some g.p.m. Q. A simple calculation now reveals that $F(S^*S)F^* \in D_n$ and hence $S^*S \in C_n$. ▲▲▲

REMARK 3.6.1. If S^*S is a circulant, then $S^{-1}PS$ need not be a circulant. It can be easily verified by taking $S = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$, say.

LEMMA 3.6.2. Let $S \in M_n$ be nonsingular. Then $S^*CS \in C_n$ for every $C \in C_n$ iff $S^{-1}PS \in C_n$.

Proof. If $S^{-1}PS \in C_n$ then by Lemma 3.6.1, $S^*S \in C_n$. Hence $S^*P^jS \in C_n$ for $j=0,1,\ldots,n-1$ which shows that $S^*(\sum_{j=0}^{n-1} c_jP^j)S \in C_n$ for any choice of $c_0,c_1,\ldots,c_{n-1}$. This, by (i)-(ii) of Theorem 3.6.1, proves the "if" part of the lemma. To prove the "only if part" we choose C=I and C=P to get $S^{-1}PS = (S^*S)^{-1}(S^*PS) \in C_n$. ▲▲▲

LEMMA 3.6.3. Let $S \in M_n$ be nonsingular. Then $S^{-1}PS \in C_n$ iff $S = F^*QF$ for some g.p.m. Q.

Proof. The necessity part follows along the proof of Lemma 3.6.1. Conversely let $S = F^*QF$ for some g.p.m. Q. Then $F(S^{-1}PS)F^* = F(F^*Q^{-1}F)P(F^*QF)F^* = Q^{-1}\tilde{D}Q \in D_n$ since $\tilde{D} = FPF^* \in D_n$ and a g.p.m. can always be expressed as a product of a diagonal matrix and a permutation matrix. ▲▲▲

Combining Lemmas 3.6.2 and 3.6.3 we obtain the following characterization.

THEOREM 3.6.2. Let $S \in M_n$ be nonsingular. Then $S^*CS \in C_n$ for every $C \in C_n$ iff $S = F^*QF$ for some g.p.m. Q.

REMARK 3.6.2. It may be noted that if any two of the matrices S^*S, S^*PS and $S^{-1}PS$ are in C_n then the remaining one is also in C_n and $S^*CS \in C_n$ for every $C \in C_n$.

We shall now consider the problem of determining the structure of any linear transformation T of M_n into M_n mapping circulants into themselves and preserving inertia of each circulant. In this connection we will require some lemmas.

LEMMA 3.6.4. Let $U: D_n \to D_n$ be a linear transformation satisfying $In(U(D)) = In(D)$ for all $D \in D_n$. Then there exists a nonnegative g.p.m. Q such that $U(D) = QDQ^*$ for all $D \in D_n$.

Proof. Let us consider the linear map $V: D_n \to \mathbb{C}^n$ defined by $V(D) = De$. Clearly V is one-to-one and onto and the inverse transformation $V^{-1}: \mathbb{C}^n \to D_n$ is defined by $V^{-1}(x) = diag(x)$ where $diag(x)$ for $x = (x_1,\ldots,x_n)^T$ is defined as $diag(x_1,\ldots,x_n)$. Define $L: \mathbb{C}^n \to \mathbb{C}^n$ by $L(x) = VUV^{-1}(x)$. L is linear. Moreover, $In(Lx) = In(V^{-1}(Lx)) = In(UV^{-1}(x)) = In(V^{-1}(x)) = In(x)$ for all $x \in \mathbb{C}^n$. Hence by Theorem 3.5.2, there exists a nonnegative g.p.m. M such that $Lx = Mx$ for all $x \in \mathbb{C}^n$. Given a diagonal matrix D, we can choose $x \in \mathbb{C}^n$, such that $D = V^{-1}(x)$. Thus $U(D) = V^{-1}LV(D) = V^{-1}(MDe) = diag(MDe)$. It is not difficult to see that $diag(\tilde{Q}\tilde{D}e) = \tilde{Q}\tilde{D}\tilde{Q}^*$ for a permutation matrix $\tilde{Q}$ and a diagonal matrix $\tilde{D}$. Since M can be expressed in the form Q_1D_1 where Q_1 is a permutation matrix and D_1 is diagonal, it follows that $U(D) = diag(Q_1D_1De) = Q_1(D_1D)Q_1^* = QDQ^*$ where $Q = Q_1D_1^{1/2}$, $D_1^{1/2}$ being the positive definite square root of D_1. This completes the proof. ▲▲▲

LEMMA 3.6.5. Let $S: C_n \to C_n$ be a linear transformation satisfying $In(S(C)) = In(C)$ for all $C \in C_n$. Then there exists a nonnegative g.p.m. Q such that $S(C) = F^*QFCF^*Q^*F$ for all $C \in C_n$, F being the DFT matrix.

Proof. Define $W: C_n \to D_n$ by $W(C) = FCF^*$ for all $C \in C_n$. Obviously W is a one-to-one linear transformation from C_n onto D_n and $W^{-1}(D) = F^*DF$ for all $D \in D_n$. Clearly $U: D_n \to D_n$ defined by $U(D) = WSW^{-1}(D)$ satisfies the hypothesis of the preceding lemma. Consequently, for all $C \in C_n$ we have $S(C) = W^{-1}UW(C) = W^{-1}U(FCF^*) = W^{-1}(QFCF^*Q^*) = F^*QFCF^*Q^*F$ where Q is anonnegative g.p.m. The proof of the lemma is complete. ▲▲▲

LEMMA 3.6.6. Let $S: M_n \to M_n$ be a linear transformation. Then S annihilates each n x n circulant, i.e., $S(C) = 0$ for all $C \in C_n$ iff there exists a linear operator L on M_n such that

$$S(A) = L((FAF^*)*(J-I)) \text{ for all } A \in M_n, \qquad (3.6.2)$$

J being the matrix whose elements are all 1 and $*$ (other than in the superscript) denoting the Hadamard product.

Proof. Since $FCF^* \in D_n$ for all $C \in C_n$, the "if" part is clear. To prove the converse, due to the linearity of S we have

$$S(A) = S(F^*((FAF^*)_*(J-I))F) + S(F^*((FAF^*)*I)F). \qquad (3.6.3)$$

The second term on the right hand side of (3.6.3) vanishes since $S(C) = 0$ for all $C \in C_n$. Hence the result follows by choosing L as the linear operator defined by

$$L(X) = S(F^*XF) \text{ for all } X \in M_n. \qquad \blacktriangle\blacktriangle\blacktriangle$$

We are now in a position to prove our main result concerning the structure of inertia-preserving linear transformations on circulants.

THEOREM 3.6.3. Let $T: M_n \to M_n$ be a linear transformation. Then for all $C \in C_n$, $T(C) \in C_n$ and $In(T(C)) = In(C)$ iff there exist a nonnegative generalized permutation matrix Q and a linear operator L on M_n such that

$$T(A) = R(A) + S(A) \text{ for all } A \in M_n \qquad (3.6.4)$$

where

$$R(A) = F^*QFAF^*Q^*F$$

and

$$S(A) = L((FAF^*)*(J-I)).$$

Proof. To prove the sufficiency we note from Lemma 3.6.6 that $S(C) = 0$ implying $T(C) = R(C)$ for all $C \in C_n$. By using the characterization mentioned in Theorem 3.6.1(vi) it is easy to see that $R(C) \in C_n$ for any circulant C. Now by applying Corollary 2.4.3, we have $In(R(C)) = In(C)$. Hence $T(C) \in C_n$ and $In(T(C)) = In(C)$ for all $C \in C_n$.

To prove the necessity, we infer from Lemma 3.6.5 that there exists a nonnegative g.p.m. Q such that $T(C) = R(C)$ for all $C \in C_n$. The transformation $\tilde{T}: M_n \to M_n$ defined by $\tilde{T}(A) = T(A) - R(A)$ is linear and annihilates circulants. Hence Lemma 3.6.6 is applicable and we have our main theorem. ▲▲▲

REMARK 3.6.3. By virtue of Corollary 3.5.1, it follows that T will have the same structure if it preserves angularity instead of inertia in the formulation given in the above theorem. The same conclusion can be drawn also from Theorem 3.2.3 in the light of the remark made just after that theorem, because C_n is closed under scalar multiplication.

4. ITERATIVE SOLUTIONS OF THE LYAPUNOV AND SYLVESTER EQUATIONS

4.1. Introduction

In this chapter we discuss some iterative methods for solving the Lyapunov matrix equation and the more general Sylvester matrix equation. The methods considered for the Lyapunov matrix equation are motivated by the classical Newton's method for determining the sign function of a matrix A with $\delta(A)=0$. It has been discussed in Section 1.2 that in the Newton method, one generates a sequence of matrices $\{A_k\}$ defined through

$$A_{k+1} = (A_k + A_k^{-1})/2 \tag{4.1.1}$$

with $A_0=A$. Here if A is positive stable, $A_k \to I$ as $k \to \infty$. In other words, for solving the Lyapunov matrix equation

$$AX + XA^* = C, \tag{4.1.2}$$

where A is positive stable, if we define

$$C_{k+1} = \alpha_k C_k + \beta_k A_k^{-1} C_k A_k^{-*}, \quad C_0=C \tag{4.1.3}$$

where

$$A_{k+1} = \alpha_k A_k + \beta_k A_k^{-1}, \quad A_0=A \tag{4.1.4}$$

and

$$\alpha_k = \beta_k = 1/2 \tag{4.1.5}$$

then, inductively, we have

$$A_k X + XA_k^* = C_k \tag{4.1.6}$$

and it follows that

$$X = (\lim_{k \to \infty} C_k)/2 \tag{4.1.7}$$

is the solution of the system (4.1.2). In the above

definition of A_{k+1}, each of the matrices A_k and A_k^{-1} are given the equal weight 1/2. In this context, as described in Section 1.4 of the survey chapter, Hoskins, Meek and Walton [205,207,208] and Barraud [163] developed several other choices of α_k and β_k to improve the performance of the method.

In Section 4.2 of this chapter we reconsider a choice of α_k and β_k given by Hoskins, Meek and Walton [207]. Several counterexamples suggest that the algorithm requires some modifications. Even with these modifications it is shown that the method does not converge for all stable (positive stable) matrices A. For the modified method, however, we are able to establish the theoretical convergence of the algorithm, in case A or -A is stable with a real spectrum. In Section 4.3 we develop a new choice of α_k and β_k and establish the theoretical convergence of the method so obtained for the class of normal matrices A. In the last two sections of this chapter we study the feasibility of the Kaczmarz and the residual projection methods for solving the Sylvester equation AX+XB=C. There our main object is to show how these projection methods could be compactly implemented for solving the Sylvester equation without giving rise to significant storage problems. Also, our numerical experience has shown that in certain cases the implementation of these methods is rather fast.

4.2. Convergence of an Algorithm of Hoskins, Meek and Walton

The key idea behind the class of iterative methods generalizing Newton's method for solving the Lyapunov matrix equation (4.1.2) is to choose α_k and β_k so that the iterative scheme (4.1.4) converges to I (or -I) more effectively when A is positive (or negative) stable. In the choice of α_k, β_k suggested by Hoskins, Meek and Walton as described in (1.4.49)-(1.4.52), for solving AX+XB=C where A and B are stable, if we specialize $B=A^T$ or A^* we get the choice

$$\alpha_k = \alpha(A_k) \text{ and } \beta_k = \beta(A_k) \tag{4.2.1}$$

in the iterative scheme (4.1.4) for a stable matrix A, where $\alpha(X)$ and $\beta(X)$ for $X \in M_n$ are defined by

$$\alpha(X) = \tilde{\alpha}(X)/\gamma(X) \tag{4.2.2}$$

$$\beta(X) = \tilde{\beta}(X)/\gamma(X) \tag{4.2.3}$$

$$\tilde{\alpha}(X) = \mathrm{tr}(X)\mathrm{tr}(X^{-2}) - n\mathrm{tr}(X^{-1}) \tag{4.2.4}$$

$$\tilde{\beta}(X) = \mathrm{tr}(X^{-1})\mathrm{tr}(X^{2}) - n\mathrm{tr}(X) \tag{4.2.5}$$

and

$$\gamma(X) = \mathrm{tr}(X^2)\mathrm{tr}(X^{-2}) - n^2. \tag{4.2.6}$$

Regarding the above choice the following may be observed:

(a) Referring to Walton [274], Hoskins et al. [207] state that

$$\lim_{k \to \infty} A_k = -I. \tag{4.2.7}$$

However, if we consider $A = \begin{bmatrix} -0.5 & 0 \\ 0 & -2 \end{bmatrix}$, then we have $\alpha_o = \beta_o = -0.4$ and thus $A_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Hence it appears that A_k tends to I rather than to -I as $k \to \infty$.

(b) Even if this probable error in the sign with I is ratified in (4.2.7), still we find that the result is not valid for a general stable matrix A. For example, if we consider

$$A = \begin{bmatrix} -1 & 4-\sqrt{13} & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & -4 & \sqrt{13}-1 \\ 0 & 0 & -1 & -2 \end{bmatrix}$$

then straightforward theoretical calculations show that the eigenvalues of A are $-1\pm i(4-\sqrt{13})^{1/2}$ and $-3\pm i(\sqrt{13}-2)^{1/2}$. Also $tr(A) = -8$, $tr(A^2) = 16$, $tr(A^{-1}) = -2$ and $tr(A^{-2}) = 2(7-\sqrt{13})/9$. Thus $\alpha_o = -1/2$, $\beta_o = 0$ and $\alpha_k = 1$, $\beta_k = 0$ for $k=1,2,\ldots\ldots$ establishing that

$$\lim_{k \to \infty} A_k = -A/2 \qquad (4.2.8)$$

which is neither $-I$ nor I.

(c) Moreover, it is important to note that the coefficients α_k and β_k given by (4.2.1) are not well defined in all cases. For, there may arise a situation when $\gamma(A_k)=0$ so that α_k and β_k are to be redefined. Two such non-trivial examples for which $\gamma(A_k)$ vanishes in the execution of the algorithm using exact arithmetic, respectively, for k=0 and k=2 are as follows:

$$A = \begin{bmatrix} \sqrt{3}-2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & 2 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} -0.5 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 \\ 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

Keeping the above observations in mind, we propose the following modification in the above choice of α_k and β_k:

$$\alpha_k = \alpha(A_k),\ \beta_k = \beta(A_k) \text{ if } \gamma(A_k) \neq 0 \qquad (4.2.9)$$

and

$$\alpha_k = n/2\mathrm{tr}(A_k),\ \beta_k = \mathrm{tr}(A_k)/2n \text{ if } \gamma(A_k)=0. \qquad (4.2.10)$$

In fact, only after developing the above choice independently, we came to know the choice of α_k and β_k defined through (1.4.49)-(1.4.52) in Hoskins et al.[207]. However, in [207] it is not explained how this choice was obtained. In this section first we explain how we developed the choice of α_k and β_k given in (4.2.9)-(4.2.10) and then we present a theoretical proof of the convergence of A_k to I in the iterative scheme defined by (4.1.4) with the choice of α_k, β_k given in (4.2.9)-(4.2.10), when the eigenvalues of $A \in M_n$ are all positive or all negative.

Now we introduce some notation to be used in this section and the following one. We denote the class of all positive (negative) stable matrices in M_n by S_n^+ (S_n^-). By SR_n^+ (SR_n^-) we mean the set of all positive (negative) stable matrices in M_n with real spectrum. $\tilde{S}_n$ denotes the set of all n x n matrices with n repeated eigenvalues and $\tilde{S}_n^{(\lambda)}$ will refer to the set of all n x n matrices all whose eigenvalues are λ. $\tilde{SR}_n^+$ refers to $\tilde{S}_n \cap SR_n^+$. Similarly $\tilde{SR}_n^-$ may be defined. In all the analysis the sequences are defined on $\mathbb{N}$, the set of nonnegative integers.

The criterion for the choice of α_k, β_k given in (4.2.9) is the minimization of the error functional $e(A_{k+1})$ defined by

$$e(Y) = tr\{(Y-I)^2\},\ Y \in M_n. \tag{4.2.11}$$

The minimization of $e(Y)$ is intended to take care of the convergence of the diagonal part of A_k to I in the iterative scheme (4.1.4). Since

$$e(A_{k+1}) = tr\{(\alpha_k A_k + \beta_k A_k^{-1} - I)^2\}, \tag{4.2.12}$$

evidently the two normal equations involved in the minimization process are

$$\alpha_k tr(A_k^2) + \beta_k n = tr(A_k) \tag{4.2.13}$$

$$\alpha_k n + \beta_k tr(A_k^{-2}) = tr(A_k^{-1}). \tag{4.2.14}$$

The above system possesses a unique solution iff $\gamma(A_k) \neq 0$, and the unique solution is given by (4.2.9). Whereas if $\gamma(A_k)=0$, the two normal equations represent one and the same equation and hence α_k and β_k can be chosen in infinitely many ways. However, for the sake of definiteness, we choose α_k and β_k as in (4.2.10). As our interest is to prove the convergence of (4.1.4) with the choice of α_k, β_k as given in (4.2.9)-(4.2.10), in the rest of this section it is assumed that $A \in SR_n^+$. (The convergence in the case $A \in SR_n^-$ will easily follow from the corresponding result for $A \in SR_n^+$.) It may be verified, in view of Lemmas 4.2.1 and 4.2.2 to follow, that the choice of α_k, β_k given in (4.2.10) satisfies both the normal equations.

Now we shall list some observations of theoretical interest

related to the iterative scheme given by (4.1.4), (4.2.9) and (4.2.10).

LEMMA 4.2.1. Let $A \in SR_n^+$. Then $\tilde{\alpha}(A) \geq 0$, $\tilde{\beta}(A) \geq 0$ and $\gamma(A) \geq 0$. Moreover, the equality holds in each case iff $A \in \tilde{S}R_n^+$.

Proof. By applying the power mean inequality [16, p.105] it can be easily shown that

$$\frac{tr(A^{-2})}{tr(A^{-1})} \geq \frac{tr(A^{-1})}{n} \geq \frac{n}{tr(A)} \geq \frac{tr(A)}{tr(A^2)} \qquad (4.2.15)$$

with the equality holding everywhere iff $A \in \tilde{S}R_n^+$ which completes the proof. ▲▲▲

LEMMA 4.2.2. Let $A \in SR_n^+$. Then in the iterative scheme defined by (4.1.4), (4.2.9) and (4.2.10), $A_k \in SR_n^+$, for each $k \in \mathbb{N}$.

Proof. The proof is by induction. Suppose $A_k \in SR_n^+ \backslash \tilde{S}_n$. Then by the previous lemma $\alpha_k > 0$ and $\beta_k > 0$ and hence $A_{k+1} \in SR_n^+$. On the other hand, if $A_k \in \tilde{S}R_n^+$, then $\gamma(A_k)=0$ and hence $\lambda_i(A_{k+1}) = (n\lambda_i(A_k)/2tr(A_k)) + (tr(A_k)/2n\lambda_i(A_k)) = 1$ for $i=1,\ldots,n$ where $\lambda_1(Y),\ldots,\lambda_n(Y)$ are the eigenvalues of Y. Thus we have shown that if $A_k \in SR_n^+$ then $A_{k+1} \in SR_n^+$. Since $A_o=A \in SR_n^+$ the lemma follows. ▲▲▲

The following lemma will be useful in proving our main theorem.

LEMMA 4.2.3. If $A \in SR_n^+ \backslash \tilde{S}_n$ and $\tilde{A} = \alpha A + \beta A^{-1}$ where $\alpha = \alpha(A)$ and $\beta = \beta(A)$, then $e(\tilde{A})$ is a continuous function of the eigenvalues of A.

Proof. Since $A \in SR_n^+ \setminus \tilde{S}_n$, we have $\gamma(A) \neq 0$ and certainly α and β are continuous functions of the eigenvalues of A. Hence $e(\tilde{A}) = tr\{(\tilde{A}-I)^2\} = \sum_{i=1}^{n} (\lambda_i(\tilde{A})-1)^2 = \sum_{i=1}^{n} \{\alpha\lambda_i(A)+(\beta/\lambda_i(A))-1\}^2$ is a continuous function of the eigenvalues of A. This completes the proof. ▲▲▲

It is convenient to denote $e(A_k)$ by e_k. It may be observed from the way in which α_k and β_k are obtained that $\{e_k\}$ is a monotonic nonincreasing sequence bounded below by zero and hence it converges to a nonnegative number. The following lemma is a consequence of this fact and the two normal equations (4.2.13) and (4.2.14).

LEMMA 4.2.4. The iterative scheme defined by (4.1.4), (4.2.9) and (4.2.10) has the following properties.

(i) $tr(A_kA_{k+1}) = tr(A_k)$ for all $k \in \mathbb{N}$.

(ii) $tr(A_k^{-1}A_{k+1}) = tr(A_k^{-1})$ for all $k \in \mathbb{N}$.

(iii) $e_k = tr(I-A_k)$ for $k=1,2,\ldots$.

(iv) $e_1 \leq n$.

(v) $tr(A_k^2) = tr(A_k)$ for $k=1,2,\ldots$.

(vi) $tr(A_1) \leq tr(A_2) \leq \ldots \leq tr(A_k) \leq \ldots \leq n$.

(vii) $tr(A_k^{-2}) \geq tr(A_k^{-1}) \geq n$ for $k=1,2,\ldots$.

(viii) $\alpha_k+\beta_k \leq 1$ for $k=1,2,\ldots$.

(ix) $\alpha_k\beta_k \leq 1/4$ for all $k \in \mathbb{N}$.

(x) $\alpha_k+\beta_k \geq \{tr(A_k)\}/n$ for $k=1,2,\ldots$.

Proof. (i) Since $A_{k+1} = \alpha_k A_k + \beta_k A_k^{-1}$ we have $tr(A_k A_{k+1}) = \alpha_k tr(A_k^2) + \beta_k n = tr(A_k)$.

(ii) Its proof is similar to that of (i).

(iii) We know $e_{k+1} = tr\{(\alpha_k A_k + \beta_k A_k^{-1} - I)^2\}$. By expanding the right hand side and simplifying with the help of normal equations, we arrive at $e_{k+1} = tr(I - A_{k+1})$ for each $k \in \mathbb{N}$, i.e., $e_k = tr(I - A_k)$ for $k=1,2,\ldots$.

(iv) $e_1 = tr\{(\alpha_o A_o + \beta_o A_o^{-1} - I)^2\} \leq tr\{(-I)^2\} = n$, since α_o, β_o minimize $e(A_1)$.

(v) Since $tr\{(A_k - I)^2\} = e_k = tr(I - A_k)$ for $k=1,2,\ldots$, we have $tr(A_k^2) = tr(A_k)$ for $k=1,2,\ldots$.

(vi) It is a consequence of the monotonicity of $\{e_k\}$ and (iii).

(vii) Since $\tilde{\alpha}(A_k) \geq 0$ and $tr(A_k) \leq n$ we conclude that $tr(A_k^{-2}) \geq tr(A_k^{-1})$. From $\tilde{\beta}(A_k) \geq 0$ and (v) we have $tr(A_k^{-1}) \geq n$.

(viii) It follows immediately from the first normal equation by using (vi).

(ix) From the second normal equation, by the well-known arithmetic-geometric mean inequality [16, p.106], we have $\{tr(A_k^{-1})\}^2 \geq 4n\alpha_k\beta_k tr(A_k^{-2})$. But $n\,tr(A_k^{-2}) \geq \{tr(A_k^{-1})\}^2$ (refer Lemma 4.2.1) for each $k \in \mathbb{N}$. This proves (ix).

(x) From the first normal equation and (v) we have $tr(A_k) = \alpha_k tr(A_k) + \beta_k n$. Now by making use of the fact that $tr(A_k) \leq n$, the result follows.

The proof of the lemma is now complete. ▲▲▲

In view of the above properties we have

LEMMA 4.2.5. For k=1,2,... the following conditions, in connection with (4.1.4), (4.2.9) and (4.2.10), are all equivalent to one another:

(i) $e_k = 0$.

(ii) $e_{k+1} = e_k$.

(iii) $tr(A_{k+1}) = tr(A_k)$.

(iv) $A_k \in \tilde{S}R_n^+$.

(v) $A_k \in \tilde{S}R_n^{(1)}$.

(vi) $tr(A_k) = n$.

(vii) $tr(A_k^2) = n$.

(viii) $tr(A_k^{-1}) = n$.

(ix) $tr(A_k^{-2}) = n$.

(x) $tr(A_k) = tr(A_k^{-1})$.

(xi) $tr(A_k^{-2}) = tr(A_k^{-1})$.

(xii) $tr(A_k A_{k+1}) = n$.

(xiii) $tr(A_k^{-1} A_{k+1}) = n$.

(xiv) $\alpha_k + \beta_k = 1$.

Proof. (vi) <==> (vii), (vi) <==> (xii) and (viii) <==> (xiii) are all obvious.

(v) ==> (iv) trivially. Conversely if (iv) holds then using $tr(A_k^2) = tr(A_k)$ we find that $A_k \in \tilde{S}R_n^{(1)}$. Thus (iv) <==> (v).

It is clear that (vi) ==> (i) ==> (v) ==> each one of (viii) to (xi). In view of $tr(A_k^{-2}) \geq tr(A_k^{-1}) \geq n$, (ix) ==> (viii). From the inequalities $tr(A_k)tr(A_k^{-1}) \geq n^2$ and $tr(A_k) \leq n$, (viii) ==> (vi) and (x) ==> (vi). Because of

$\tilde{\alpha}(A_k) \geq 0$ and $tr(A_k) \leq n$, (xi) ==> (vi). Thus we have shown that (i) and (iv) to (xiii) are all equivalent to one another. Moreover (ii) <==> (iii) is immediate.

To complete the proof it suffices to show that (ii) <==> (iv) and (vi) <==> (xiv). If $e_{k+1} = e_k$, then $\alpha_k = 1$, $\beta_k = 0$ is a choice of the parameters. Hence A_k should belong to $\tilde{SR}_n^+$ (refer the proof of Lemma 4.2.1). Hence (ii) ==> (iv) ==> (v) ==> (ii). In view of the first normal equation, it easily follows that (vi) ==> (xiv). On the other hand, if $\alpha_k+\beta_k = 1$, then again from the first normal equation we get $\alpha_k = 1$ or $tr(A_k) = n$. If $\alpha_k = 1$ then $\beta_k = 0$ which in turn implies that $A_k \in \tilde{SR}_n^+$ and hence (vi). This completes the proof of the lemma. ▲▲▲

An immediate consequence of the above lemma is

COROLLARY 4.2.1. For k=1,2,..., if $A_k \in SR_n^+ \setminus \tilde{S}_n$ then

(i) $e_{k+1} < e_k$.

(ii) $tr(A_k) < n < tr(A_k^{-1}) < tr(A_k^{-2})$.

(iii) $0 < \alpha_k+\beta_k < 1$.

Next we prove another simple lemma which will play a role in the proof of the main theorem.

LEMMA 4.2.6. If for $k=k_o$, $A_k \in \tilde{SR}_n^+$ then $\alpha_k = \beta_k = 1/2$ for all $k>k_o$. In fact, it holds even for $k=k_o$ except possibly when $k_o=0$.

Proof. From Lemma 4.2.5, we have $A_k \in \tilde{SR}_n^+$ ==> $A_k \in \tilde{SR}_n^{(1)}$ for k=1,2,... implying that $\gamma(A_k) = 0$. Hence by using (4.2.10),

we get $\alpha_k = \beta_k = 1/2$. Moreover, if $A_k \in \tilde{S}R_n^{(1)}$ for $k=k_o$ then it holds for $k > k_o$ also. This completes the proof.▲▲▲

Finally, we state one lemma due to Laasonen [297] which is an essential tool used in the proof of the main convergence theorem.

LEMMA 4.2.7. Let $\{a_k\}$, $\{b_k\}$ and $\{c_k\}$ denote three sequences of real numbers connected by $a_{k+1} = c_k a_k + b_k$. If $\{b_k\}$ and $\{c_k\}$ have limits b and 0, respectively, the sequence $\{a_k\}$ converges to the limit b.

Now we prove the following main convergence theorem related to our modified algorithm for solving the Lyapunov matrix equation (4.1.2) assuming that $A \in SR_n^+$.

THEOREM 4.2.1. Let

$$A_{k+1} = \alpha_k A_k + \beta_k A_k^{-1} \text{ for } k \in \mathbb{N}$$

with $A_o \in SR_n^+$, where

$$\alpha_k = \alpha(A_k),\ \beta_k = \beta(A_k) \text{ if } A_k \in SR_n^+ \setminus \tilde{S}_n \text{ (i.e. if } \gamma(A_k) \neq 0)$$

and

$$\alpha_k = n/2\mathrm{tr}(A_k),\ \beta_k = \mathrm{tr}(A_k)/2n \text{ if } A_k \in \tilde{S}R_n^+ \text{ (i.e. if } \gamma(A_k)=0).$$

Then as $k \to \infty$, $\alpha_k \to 1/2$, $\beta_k \to 1/2$ and $A_k \to I$.

Proof. We shall establish this result in four stages.

Stage 1. Suppose for some k, say k_o, $A_k \in \tilde{S}R_n^+$. Then the proof is very easy. By Lemma 4.2.6, $\alpha_k = \beta_k = 1/2$ for all $k > k_o$ and hence it remains to prove only that $A_k \to I$. In order to prove this, it is sufficient to establish the

convergence of the classical iterative scheme

$$A_{k+1} = (A_k + A_k^{-1})/2,\ A_o \in SR_n^+. \qquad (4.2.16)$$

Hoskins, Meek and Walton have given a proof for this in [205]. However, the following direct, alternative proof may be given. From (4.2.16), by a simple calculation we have

$$(A_{k+1}-I)(A_{k+1}+I)^{-1} = \{(A_k-I)(A_k+I)^{-1}\}^2. \qquad (4.2.17)$$

If we put $(A_k-I)(A_k+I)^{-1}$ as E_k, then it follows that

$$E_k = E_o^{2^k}. \qquad (4.2.18)$$

Since $A_o \in SR_n^+$, $\rho(E_o) < 1$ and hence

$$\lim_{k \to \infty} E_k = 0. \qquad (4.2.19)$$

As $A_k = (I-E_k)^{-1}(I+E_k)$, clearly $A_k \to I$ as $k \to \infty$.

We remark that the above proof holds even if $A_o \in S_n^+$.

Stage 2. We shall now consider the case when for all $k \in \mathbb{N}$, $A_k \in SR_n^+ \backslash \tilde{S}_n$, so that $\alpha_k = \alpha(A_k)$ and $\beta_k = \beta(A_k)$ for all $k \in \mathbb{N}$. It will be convenient to express the iterative scheme (4.1.4) using Jordan canonical form of the initial matrix A_o. If $J_o = T^{-1}A_oT$ is the Jordan form of A_o, from (4.1.4), equivalently,

$$J_{k+1} = \alpha_k J_k + \beta_k J_k^{-1},\ J_k = T^{-1}A_kT \text{ for every } k \in \mathbb{N} \qquad (4.2.20)$$

with $J_o \in SR_n^+$ and $\alpha_k = \alpha(J_k)$, $\beta_k = \beta(J_k)$. Note that J_k need not be the Jordan form of A_k for $k=1,2,\ldots$. However, J_k is an upper triangular matrix having the eigenvalues of A_k along its main diagonal. In fact, the convergence of J_k to I will

imply the convergence of A_k to I and vice versa. Note that $e_k = e(A_k) = e(J_k)$. Let us prove at this stage that $e_k \to 0$ as $k \to \infty$.

The boundedness of $\{e_k\}$ indicates that each one of the n sequences $\{\lambda_i(J_k)\}$, i=1,...,n is bounded. Hence by Cantor's diagonal process there exists a subsequence $\{k_p\}$ of nonnegative integers where $k_o < k_1 < k_2 < \ldots$ such that $\{\lambda_i(J_{k_p})\} \to \lambda_i$ say, as $p \to \infty$ for each i=1,...,n. Since a sequence of positive numbers can converge only to a nonegative number, we have one of the following situations.

(i) $\lambda_i=0$ for all i=1,...,n.

(ii) $\lambda_i=0$ for at least one i and at most (n-1) indices $i \in \{1,\ldots,n\}$.

(iii) $\lambda_i > 0$ for all i=1,...,n and $\lambda_i \neq \lambda_j$ for at least one pair (i,j).

(iv) For some $\lambda > 0$, $\lambda_i = \lambda$ for all i=1,...,n.

Assume, eventually to reach a contradiction that $\lambda_i = 0$, for all i=1,...,n. Since $e_1 \leq n$, it follows that $e_2 < n$. Hence there exists some $q \in \mathbb{N}$ such that $e_{k_q} < n$ and evidently for some $c > 0$, $e_{k_q} < n-c$. Moreover, from the assumption, $\lim_{p\to\infty} e_{k_p} = n$. Therefore, it is true that for any given $c > 0$, $|e_{k_p}-n| < c/2$ holds for all sufficiently large values of p. Hence for some $r > q$, $e_{k_r} > n-(c/2)$. Since $\{e_{k_p}\}$ is a monotonic decreasing sequence, we have $e_{k_r} < e_{k_q}$ which shows that $n-(c/2) < n-c$ and we have a contradiction.

Next, suppose $S = \{i_o, i_1, \ldots, i_m\}$, where $0 \leq m < n-1$, to be the set of indices i for which λ_i vanishes. Clearly, without loss of generality, we can assume that along a subsequence $\{k_{1,p}\}$ of the subsequence $\{k_p\}$,

$$\lim_{p \to \infty} \frac{\lambda_{i_o}(J_{k_{1,p}})}{\lambda_{i_j}(J_{k_{1,p}})} = c_j, \quad c_j \in [0,\infty) \text{ for } j=0,1,\ldots,m. \tag{4.2.21}$$

To simplify the notation let us put $\lambda_{i_j}(J_{k_{1,p}}) = \tilde{\lambda}_{i_j}$. Let $\mu_1, \mu_2, \ldots, \mu_{n-m-1}$ be the nonvanishing λ_i's. Note that there is at least one nonvanishing λ_i. It is not hard to see that

$$\alpha_{k_{1,p}} \simeq \frac{(\lambda_{i_o})^{-2}(\Sigma\mu_i)(1+\Sigma c_j^2) - n(\lambda_{i_o})^{-1}(1+\Sigma c_j)}{(\lambda_{i_o})^{-2}(\Sigma\mu_i^2)(1+\Sigma c_j^2) - n^2} \tag{4.2.22}$$

where in the summations i varies from 1 to n-m-1 and j varies from 1 to m. Simplifying further, we see that

$$\alpha_{k_{1,p}} \simeq \frac{\Sigma\mu_i}{\Sigma\mu_i^2} > 0. \tag{4.2.23}$$

Hence

$$\lim_{p \to \infty} \alpha_{k_{1,p}} > 0. \tag{4.2.24}$$

Since

$$\lambda_{i_o}(J_{k_{1,p}+1}) = \alpha_{k_{1,p}}\tilde{\lambda}_{i_o} + \beta_{k_{1,p}}(\tilde{\lambda}_{i_o})^{-1} \tag{4.2.25}$$

and $\alpha_{k_{1,p}}$ is bounded we have

$$\lim_{p \to \infty} \lambda_{i_o}(J_{k_{1,p}+1}) = \lim_{p \to \infty} \beta_{k_{1,p}} (\tilde{\lambda}_{i_o})^{-1}. \qquad (4.2.26)$$

On simplification, the limit on the right hand side becomes $(1+\Sigma c_j)/(1+\Sigma c_j^2)$ which is nonzero. In particular it follows that

$$\lim_{p \to \infty} \beta_{k_{1,p}} = 0. \qquad (4.2.27)$$

If i is such that $\lambda_i \neq 0$, using (4.2.23) and (4.2.27) it follows that

$$\lim_{p \to \infty} \lambda_i(J_{k_{1,p}+1}) > 0. \qquad (4.2.28)$$

Hence the set of j's for which

$$\lim_{p \to \infty} \lambda_j(J_{k_{1,p}+1}) = 0 \qquad (4.2.29)$$

has at most (n-2) elements. Put $k_{1,p}+1 = k_p^{(1)}$. From the above analysis we conclude that along the subsequence $\{k_p^{(1)}\}$ of nonnegative integers, each one of the n sequences $\{\lambda_i(J_{k_p^{(1)}})\}$, i=1,...,n converges to nonnegative numbers and among them at most (n-2) limits are zero. Continue the above process of constructing subsequences along which each time the number of zero limits is decreasing at least by one. Hence after at most (n-1) such constructions, we will have a subsequence, call it $\{s_p\}$, of nonnegative integers where $s_o < s_1 < s_2 < \ldots$ such that $\{\lambda_i(J_{s_p})\}$ converges to a positive number say d_i, for each i=1,...,n.

Suppose $d_i \neq d_j$ for at least one pair (i,j) and let $D = \text{diag}(d_1,\ldots,d_n)$. If for each $k \in \mathbb{N}$, D_k denotes the

diagonal matrix having the same leading diagonal as that of J_k, obviously

$$\lim_{p \to \infty} D_{s_p} = D. \tag{4.2.30}$$

Let $\{e_k\}$ converge to e. Now,

$$e = \lim_{p\to\infty} e_{s_p} = \lim_{p\to\infty} \sum_{i=1}^{n} \{\lambda_i(J_{s_p})-1\}^2 = \sum_{i=1}^{n} (d_i-1)^2 = e(D). \tag{4.2.31}$$

On the other hand, if we denote $e(D_{s_p+1})$ by $e_1(D_{s_p})$, then

$$e = \lim_{p\to\infty} e_{s_p+1} = \lim_{p\to\infty} e(D_{s_p+1}) = \lim_{p\to\infty} e_1(D_{s_p}). \tag{4.2.32}$$

Upon using the continuity result exhibited in Lemma 4.2.3, the last mentioned limit becomes $e_1(D)$ which is same as $e(\alpha D+\beta D^{-1})$ where $\alpha = \alpha(D)$ and $\beta = \beta(D)$. Since $D \in SR_n^+ \setminus \tilde{S}_n$, we know that $e(\alpha D+\beta D^{-1}) < e(D)$. Thus in view of (4.2.31) we have a contradiction.

Still we are left with the case that $d_1 = d_2 = \ldots = d_n = d$ say. Then $\lambda_i(J_{s_p}) \to d$ for every i=1,...,n and hence for any given $\varepsilon>0$, there exists p_o such that $p > p_o$ implies $|\lambda_i(J_{s_p})-d| < \varepsilon$ for every i=1,...,n. Therefore, $\lambda_i(J_{s_p}) = d + \varepsilon\theta_{i,s_p}$ for all $p > p_o$ where each $\theta_{i,s_p} \in (-1,1)$. Since

$$e_{s_p+1} = \sum_{i=1}^{n} \{\alpha_{s_p} \lambda_i(J_{s_p}) + \beta_{s_p} (\lambda_i(J_{s_p}))^{-1} - 1\}^2 \tag{4.2.33}$$

where α_{s_p} and β_{s_p} are the unique optimal values for e_{s_p+1} to be minimum, it follows that

$$e_{s_p+1} < \sum_{i=1}^{n} \{\tfrac{1}{d}\lambda_i(J_{s_p})-1\}^2. \tag{4.2.34}$$

it follows that $\lim_{k \to \infty} \alpha_k = 1/2$. By similar computations it can be shown that $\beta_k \to 1/2$.

It may be noted that since $D_k \to I$, $tr(A_k) = tr(D_k) \to n$ and $tr(A_k^2) = tr(D_k^2) \to n$. Hence once we prove that $\alpha_k \to 1/2$ then from the first normal equation also it follows that $\beta_k \to 1/2$.

Stage 4. Now we have all the required material to complete the proof of our main theorem. To prove $J_k \to I$, we follow a procedure used by Laasonen [297] for Newton's method. If $J_{o,i}$, $i=1,\ldots,r$ are the elementary Jordan blocks of J_o so that $J_o = \sum_{i=1}^{r} \oplus J_{o,i}$, where $\oplus$ denotes the direct sum, then we have

$$J_{k+1,i} = \alpha_k J_{k,i} + \beta_k J_{k,i}^{-1}, \quad i=1,\ldots,r \tag{4.2.36}$$

where $\alpha_k = \alpha(J_k)$, $\beta_k = \beta(J_k)$ and $J_k = \sum_{i=1}^{r} \oplus J_{k,i}$. It should be noted that $J_{k,i}$ is an upper triangular Toeplitz matrix [293] for $k \in \mathbb{N}$, $i=1,\ldots,r$. Let for a fixed i,

$$J_{k,i} = \begin{bmatrix} d_o^{(k)} & d_1^{(k)} & \ldots & d_{m-2}^{(k)} & d_{m-1}^{(k)} \\ 0 & d_o^{(k)} & \ldots & d_{m-3}^{(k)} & d_{m-2}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & d_o^{(k)} & d_1^{(k)} \\ 0 & 0 & \ldots & 0 & d_o^{(k)} \end{bmatrix}$$

which may be denoted by $[d_o^{(k)}, d_1^{(k)}, \ldots, d_{m-1}^{(k)}]$ for convenience. From (4.2.36) we have

$$J_{k+1,i}\, J_{k,i} = \alpha_k J_{k,i}^2 + \beta_k I_m. \tag{4.2.37}$$

By performing the matrix multiplications and equating the corresponding elements on both sides we get

$$d_o^{(k)} d_o^{(k+1)} = \alpha_k d_o^{(k)} d_o^{(k)} + \beta_k \tag{4.2.38}$$

and

$$d_o^{(k)} d_s^{(k+1)} + d_1^{(k)} d_{s-1}^{(k+1)} + \ldots + d_s^{(k)} d_o^{(k+1)}$$
$$= \alpha_k \{ d_o^{(k)} d_s^{(k)} + d_1^{(k)} d_{s-1}^{(k)} + \ldots + d_s^{(k)} d_o^{(k)} \}$$
$$\text{for } s=1,\ldots,m-1. \tag{4.2.39}$$

Equation (4.2.38) is consistent with the facts $\alpha_k \to 1/2$ $\beta_k \to 1/2$ and $d_o^{(k)} \to 1$. Solving (4.2.39) for $d_s^{(k+1)}$ we find

$$d_s^{(k+1)} = \{2\alpha_k - (d_o^{(k+1)}/d_o^{(k)})\} d_s^{(k)} + (1/d_o^{(k)}) P \tag{4.2.40}$$

where P is a quadratic polynomial in $d_1^{(k)}, \ldots, d_{s-1}^{(k)}, d_1^{(k+1)}, \ldots, d_{s-1}^{(k+1)}$. If $d_1^{(k)}, d_2^{(k)}, \ldots, d_{s-1}^{(k)} \to 0$ then by Lemma 4.2.7, $d_s^{(k)} \to 0$. Writing the equation corresponding to s=1, from (4.2.39) we have

$$d_1^{(k+1)} = \{2\alpha_k - (d_o^{(k+1)}/d_o^{(k)})\} d_1^{(k)}. \tag{4.2.41}$$

Again by Lemma 4.2.7, $d_1^{(k)} \to 0$. Hence by induction, we have

$$\lim_{k \to \infty} d_s^{(k)} = 0 \text{ for } s=1,2,\ldots,m-1. \tag{4.2.42}$$

Thus we have established that

$$\lim_{k \to \infty} J_{k,i} = I_m \tag{4.2.43}$$

It is true for i=1,...,r and of course m depends on i.

This completes the proof of our main theorem. ▲▲▲

REMARK 4.2.1. The result of the above theorem remains valid even if the assumption $A \in SR_n^+$ is replaced by $A \in SR_n^-$. For, as may be easily verified, in this case $A_1 \in SR_n^+$. If $A \in S_n^+$ or S_n^- then in view of the example provided in observation (b) in the beginning of this section, A_k need not converge to I in general.

REMARK 4.2.2. If A is any nonsingular 2x2 matrix, then $A_1 = I$, i.e., the convergence takes place in a single iteration. For, if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ where $ad-bc \neq 0$, it is easily seen that whether the eigenvalues λ_1 and λ_2 of A or equal or not,

$$\alpha_o = 1/(\lambda_1+\lambda_2),\ \beta_o = \lambda_1\lambda_2/(\lambda_1+\lambda_2) \qquad (4.2.44)$$

so that

$$A_1 = \frac{1}{\lambda_1+\lambda_2}\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\cdot\frac{1}{\lambda_1\lambda_2}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

because $a+d=\lambda_1+\lambda_2$.

REMARK 4.2.3. If for some k, $\gamma(A_k)$ is sufficiently small, the eigenvalues of A_k are almost equal. Therefore from next iterate onwards the method starts behaving like Newton's method, i.e., α_k and β_k come close to 1/2. In this situation there is little point in working with the formulae given in (4.2.9). Hoskins, Meek and Walton [207] suggest to switch over to the formulae given in (1.4.41)-(1.4.44). There the computations of α_k and β_k are based on some norms of A_k and A_k^{-1}. However in our simulations we have found that our choice given by (4.2.10) compares favourably with the choice

of Hoskins et al.[208] and it does not require so much computations.

Finally, we present some numerical simulations which provide a comparison of the performance of the iterative scheme (4.1.4) for the following four choices of α_k and β_k.

(1) Newton's method ($\alpha_k = \beta_k = 1/2$).

(2) The choice given in (1.4.41)-(1.4.44) due to Hoskins, Meek and Walton [208].

(3) The choice given in (1.4.45)-(1.4.46) due to Barraud [163].

(4) Our least squares choice given in (4.2.9)-(4.2.10).

In the following examples the symbols N, H, B, L denote respectively the above choices (1)-(4). We present the matrix A and the number of iterations required for these choices to have the convergence upto six decimal places. All the computations were performed on the DEC 1090 using FORTRAN IV with the single precision at the Indian Institute of Technology, Kanpur.

EXAMPLE 4.2.1.

$$A = \begin{bmatrix} 30.0 & 32.89 \\ 9.0 & 10.00 \end{bmatrix}$$

N-8, H-4, B-4, L-1.

EXAMPLE 4.2.2.

$$A = \begin{bmatrix} 2.5 & 5.5 & 2.0 \\ 2.0 & 6.0 & 2.0 \\ 1.5 & -1.5 & 2.0 \end{bmatrix}$$

N-6, H-4, B-4, L-2.

EXAMPLE 4.2.3.

$$A = \begin{bmatrix} 0.2 & 4.0 & 6.3 & -5.4 \\ & 1.4 & -6.5 & 6.6 \\ & \bigcirc & 3.2 & -0.6 \\ & & & 5.0 \end{bmatrix}$$

N-5, H-7, B-7, L-4.

EXAMPLE 4.2.4.

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

N-6, H-5, B-6, L-4.

EXAMPLE 4.2.5.

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ & 2 & 3 & 4 & 5 & 6 \\ & & 3 & 4 & 5 & 6 \\ & \bigcirc & & 4 & 5 & 6 \\ & & & & 5 & 6 \\ & & & & & 6 \end{bmatrix}$$

N-6, H-5, B-5, L-3.

EXAMPLE 4.2.6.

$$A = \begin{bmatrix} 4.5 & 2.3 & 1.4 & -1.2 & 3.6 & 1.4 \\ & 2.1 & 4.1 & 1.2 & -3.5 & 6.7 \\ & & 1.2 & 0.8 & 1.0 & -1.7 \\ & & \bigcirc & 0.8 & -2.1 & 1.5 \\ & & & & 5.1 & -1.2 \\ & & & & & 0.3 \end{bmatrix}$$

N-5, H-6, B-6, L-5.

EXAMPLE 4.2.7.

$$A = \begin{bmatrix} 6 & 5 & & \bigcirc & & \\ 5 & 5 & 4 & & & \\ 4 & 4 & 4 & 3 & & \\ 3 & 3 & 3 & 3 & 2 & \\ 2 & 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

N-7, H-5, B-6, L-4.

EXAMPLE 4.2.8.

$$A = J_{10}(1)$$

where $J_m(\lambda)$ is defined as the m x m matrix whose diagonal elements are all λ and the superdiagonal elements are all unity and the remaining elements are all zero.

N-4, H-4, B-4, L-4.

EXAMPLE 4.2.9.

$$A = J_3(1.8) + J_2(0.79) + J_1(1.02) + J_1(1.55) + J_1(2.52) + J_1(4.36) + J_1(6.21)$$

N-6, H-5, B-7, L-5.

From the above examples it is clear that the least squares choice compares favourably with the remaining three choices.

4.3. Another Iterative Method for Solving the Lyapunov Matrix Equation

In the algorithm of the previous section, the proof of the convergence required that the spectrum of A be real. In this section we develop another choice of α_k and β_k for which this restriction is not necessary. However in the case of the algorithm of this section, the assumption that the matrix A be normal will be required. Thus the method developed would be applicable for an arbitrary positive or negative stable normal matrix.

The criterion for determining α_k and β_k for the algorithm of this section is the minimization of the functional

$\|A_{k+1}-I\|_F^2 = tr\{(A_{k+1}^*-I)(A_{k+1}-I)\}$. The corresponding normal equations become

$$2\alpha_k tr(A_k^*A_k) + \beta_k tr(A_k^*A_k^{-1}+A_k^{-*}A_k) = tr(A_k^*+A_k) \qquad (4.3.1)$$

and

$$\alpha_k tr(A_k^*A_k^{-1}+A_k^{-*}A_k) + 2\beta_k tr(A_k^{-*}A_k^{-1}) = tr(A_k^{-*}+A_k^{-1}). \qquad (4.3.2)$$

In the derivation of these equations, it has been assumed that α_k and β_k are to be real. Hence if for $X \in M_n$ we define

$$\gamma(X) = 4tr(X^*X)tr(X^{-*}X^{-1}) - \{tr(X^*X^{-1}+X^{-*}X)\}^2 \qquad (4.3.3)$$

$$\alpha(X) = \tilde{\alpha}(X)/\gamma(X) \qquad (4.3.4)$$

$$\beta(X) = \tilde{\beta}(X)/\gamma(X) \qquad (4.3.5)$$

where

$$\tilde{\alpha}(X) = 2tr(X^*+X)tr(X^{-*}X^{-1}) - tr(X^*X^{-1}+X^{-*}X)tr(X^{-*}+X^{-1}) \qquad (4.3.6)$$

$$\tilde{\beta}(X) = 2tr(X^{-*}+X^{-1})tr(X^*X) - tr(X^*X^{-1}+X^{-*}X)tr(X^*+X) \qquad (4.3.7)$$

then

$$\alpha_k = \alpha(A_k) \text{ and } \beta_k = \beta(A_k) \text{ provided } \gamma(A_k) \neq 0. \qquad (4.3.8)$$

It may be noted that the expressions of $\alpha(X)$, $\beta(X)$ and $\gamma(X)$ in this section are different from those of the preceding section.

If $\gamma(A_k) = 0$, as in the previous section, put

$$\alpha_k = n/2tr(A_k) \text{ and } \beta_k = tr(A_k)/2n. \qquad (4.3.9)$$

As remarked before the convergence of the algorithm with the present choice of α_k and β_k will be proved only under the assumption that $A \in NS_n^+$ where $NS_n^+ = N_n \cap S_n^+$. If A is not normal the method may or may not be applicable as is clear from the following two examples.

EXAMPLE 4.3.1. With $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ simple calculations show that $A_1 = A_2 = \ldots = I$. Hence the method works in this case even though A is not normal.

EXAMPLE 4.3.2. With $A = \begin{bmatrix} 1 & 6 & 24 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{bmatrix}$,

$\mathrm{tr}(A^*A) = 651$, $\mathrm{tr}(A^{-*}A^{-1}) = 219$, $\mathrm{tr}(A^*A^{-1}+A^{-*}A) = 438$, $\mathrm{tr}(A^*+A) = \mathrm{tr}(A^{-*}+A^{-1}) = 6$ and hence $\alpha_o = 0$ and $\beta_o = 1/73$ It implies that $A_1 = (1/73)A^{-1}$ and consequently that $A_k = (1/73)A^{-1}$ for k=2,3,... and hence the convergence is to the matrix $(1/73)A^{-1}$ which is not equal to I.

Before stating and proving the main convergence result of this section we establish some preliminary lemmas and properties of the new algorithm.

LEMMA 4.3.1. For $A \in S_n^+$, $\gamma(A) \geq 0$ and the equality holds iff A is a positive scalar matrix.

Proof. Writing $A = (a_{rs})$ and $A^{-1} = (b_{rs})$, we have

$$\gamma(A) = 4 \Sigma |a_{rs}|^2 \Sigma |b_{rs}|^2 - \{\Sigma(\bar{a}_{rs}b_{rs}+a_{rs}\bar{b}_{rs})\}^2 \quad (4.3.10)$$

where in each summation r,s run from 1 to n. If we put

$$a_{rs} = x_{rs}\exp(i\theta_{rs}), \; b_{rs} = y_{rs}\exp(i\phi_{rs}) \quad (4.3.11)$$

where $i = \sqrt{-1}$, $x_{rs}, y_{rs} \geq 0$ and $0 \leq \theta_{rs}, \phi_{rs} < 2\pi$ then from (4.3.11)

$$\gamma(A) = 4 \Sigma x_{rs}^2 \Sigma y_{rs}^2 - 4\{\Sigma x_{rs}y_{rs}\cos(\theta_{rs}-\phi_{rs})\}^2 \quad (4.3.12)$$

and hence

$$\gamma(A) \geq 4 \Sigma x_{rs}^2 \Sigma y_{rs}^2 - 4(\Sigma x_{rs}y_{rs})^2 \quad (4.3.13)$$

with the equality if and only if either

$$\theta_{rs} = \phi_{rs} \text{ for all } r,s=1,\ldots,n \tag{4.3.14}$$

or

$$|\theta_{rs}-\phi_{rs}| = \pi \text{ for all } r,s=1,\ldots,n. \tag{4.3.15}$$

Moreover, by the well-known Cauchy's inequality

$$\Sigma x_{rs}^2 \Sigma y_{rs}^2 \geq (\Sigma x_{rs}y_{rs})^2 \tag{4.3.16}$$

with the equality iff

$$x_{rs} = \mu y_{rs}, \quad \mu > 0 \text{ for all } r,s=1,\ldots,n. \tag{4.3.17}$$

The positivity restriction on μ in (4.3.17) is due to the fact that x_{rs}, $y_{rs} \geq 0$ and $A \neq 0$. From these arguments it is clear that $\gamma(A) \geq 0$ and the equality holds iff in addition to (4.3.17) either (4.3.14) or (4.3.15) holds. In other words, $\gamma(A) = 0$ iff A^2 is a positive or negative scalar matrix. Since $A \in S_n^+$, if we use Jordan canonical form, it is not hard to see that $\gamma(A)$ vanishes iff A is a positive scalar matrix. This completes the proof. ΔΔΔ

LEMMA 4.3.2. For $A \in NS_n^+$, $\tilde{\alpha}(A) \geq 0$, $\tilde{\beta}(A) \geq 0$ and in either case the equality holds iff A is a positive scalar matrix.

Proof. Let U be the unitary diagonalizer of A so that $U^*AU = D$ where $D = \text{diag}(\lambda_1,\ldots,\lambda_n)$ such that $\text{Re}(\lambda_i) = x_i > 0$ and $\text{Im}(\lambda_i) = y_i$ for $i=1,\ldots,n$. Then

$$\tilde{\alpha}(A) = \tilde{\alpha}(D) = 4\sum_{i=1}^{n} x_i \sum_{j=1}^{n} \frac{1}{x_j^2+y_j^2} - 4\sum_{i=1}^{n} \frac{x_i^2-y_i^2}{x_i^2+y_i^2} \sum_{j=1}^{n} \frac{x_j}{x_j^2+y_j^2}$$

$$= 8\sum_{i=1}^{n} \frac{x_i y_i^2}{(x_i^2+y_i^2)^2} + 4\sum_{\substack{i,j=1 \\ i<j}}^{n} \frac{(x_i+x_j)\{(x_i-x_j)^2+(y_i^2+y_j^2)\}}{(x_i^2+y_i^2)(x_j^2+y_j^2)}.$$

Since $x_i > 0$ and y_i is real for i=1,...,n, it follows that $\tilde{\alpha}(A) \geq 0$ and the equality holds iff $x_i = x_j$ for every pair (i,j) and $y_i = 0$ for each i. This completes the proof of the lemma concerning $\tilde{\alpha}(A)$. The other part follows immediately from the fact that $\tilde{\beta}(A) = \tilde{\alpha}(A^{-1})$ and $A^{-1} \in NS_n^+$. ΔΔΔ

We now state some useful corollaries which are immediate consequences of the previous two lemmas.

COROLLARY 4.3.1. If $A \in NS_n^+$ and is not a scalar matrix, then $\gamma(A)$, $\alpha(A)$ and $\beta(A)$ are all strictly positive.

It follows from Corollary 4.3.1 that for $A \in NS_n^+$ the choice (4.3.9) for α_k and β_k in case $\gamma(A_k) = 0$ also satisfies the normal equations (4.3.1)-(4.3.2).

COROLLARY 4.3.2. Over the set of non-scalar NS_n^+ matrices A, $\alpha(A)$ and $\beta(A)$ are continuous functions of A.

COROLLARY 4.3.3. If $A \in NS_n^+$ and is not a scalar matrix and if $B = \alpha A + \beta A^{-1}$ where $\alpha = \alpha(A)$, $\beta = \beta(A)$, then $\|B-I\|_F^2$ is a continuous function of A.

LEMMA 4.3.3. In the iterative scheme defined by (4.1.4), (4.3.8) and (4.3.9) if $A_0 = A \in NS_n^+$, then $A_k \in NS_n^+$ for each $k \in \mathbb{N}$.

Proof. The proof is by induction. Suppose $A_k \in NS_n^+$. If A_k is not a scalar matrix, in view of Corollary 4.3.1, $\alpha_k, \beta_k > 0$. Since $A_{k+1} = \alpha_k A_k + \beta_k A_k^{-1}$, it is clear that $A_{k+1} \in NS_n^+$. On the other hand, if A_k is a scalar matrix then by (4.3.9), $A_{k+1} = I$ which is obviously a normal positive

stable matrix. As $A_o \in NS_n^+$, the lemma follows. ▲▲▲

LEMMA 4.3.4. Let $e_k = e(A_k) = \|A_k - I\|_F^2$. Then:

(a) $\{e_k\}$ is a monotonic nonincreasing sequence bounded below by zero and hence it converges to a nonnegative number.

(b) If, for a certain k, say k_o, A_k is a scalar matrix, then $A_k = I$ and hence $\alpha_k = \beta_k = 1/2$ for every $k > k_o$.

(c) If for every $k \in \mathbb{N}$, A_k is a non-scalar matrix then $\gamma(A_k) \neq 0$, α_k and β_k are positive and are given all through by (4.3.8). Moreover, in this case $\{e_k\}$ is a strictly decreasing sequence. Also $e_1 < n$.

This lemma plays an important role in the course of proof of our main convergence theorem. Its proof, being simple, is omitted.

Some properties of the algorithm of the present section, which are analogous to those of the previous section, are summarized as follows:

THEOREM 4.3.1. The iterative scheme defined by (4.1.4), (4.3.8) and (4.3.9) has the following properties, given $A \in NS_n^+$.

(i) $\mathrm{Re}\{\mathrm{tr}(A_k^* A_{k+1})\} = \mathrm{Re}\{\mathrm{tr}(A_k)\}$, $k \in \mathbb{N}$.

(ii) $\mathrm{Re}\{\mathrm{tr}(A_k^{-*} A_{k+1})\} = \mathrm{Re}\{\mathrm{tr}(A_k^{-1})\}$, $k \in \mathbb{N}$.

(iii) $e_k = \mathrm{Re}\{\mathrm{tr}(I - A_k)\}$, $k=1,2,\ldots$.

(iv) $\mathrm{Re}\{\mathrm{tr}(A_1)\} \leq \mathrm{Re}\{\mathrm{tr}(A_2)\} \leq \ldots \leq \mathrm{Re}\{\mathrm{tr}(A_k)\} \ldots \leq n$.

(v) $\mathrm{tr}(A_k^* A_k) = \mathrm{Re}\{\mathrm{tr}(A_k)\}$, $k=1,2,\ldots$.

(vi) $e_1 \leq n$.

(vii) $\mathrm{Re}\{\mathrm{tr}(A_k^{-1})\} \geq \mathrm{Re}\{\mathrm{tr}(A_k^* A_k^{-1})\}$, $k=1,2,\ldots$.

(viii) $\mathrm{Re}\{\mathrm{tr}(A_k^* A_k^{-1})\} \leq n$, $k \in \mathbb{N}$.

(ix) $\alpha_k + \beta_k \geq \{\mathrm{Re}(\mathrm{tr}(A_k))\}/n$, $k=1,2,\ldots$.

Proof. (i) From (4.1.4), $A_k^* A_{k+1} = \alpha_k A_k^* A_k + \beta_k A_k^* A_k^{-1}$. Considering the real part of the trace on both sides and by making use of the first normal equation (4.3.1), the result follows.

(ii) This can be shown by making use of the second normal equation.

(ii) We know $e_{k+1} = \mathrm{tr}\{(A_{k+1}^* - I)(A_{k+1} - I)\}$. By expanding the right hand side and simplifying with the help of normal equations, we arrive at $e_{k+1} = \mathrm{Re}\{\mathrm{tr}(I - A_{k+1})\}$ for each $k \in \mathbb{N}$ and from this the required result follows.

(iv) It is a consequence of the monotonicity of $\{e_k\}$ and the above result.

(v) Since $\mathrm{tr}\{(A_k^* - I)(A_k - I)\} = e_k = \mathrm{Re}\{\mathrm{tr}(I - A_k)\}$, we have (v).

(vi) $e_1 = \mathrm{tr}\{(\alpha_o A_o^* + \beta_o A_o^{-*} - I)(\alpha_o A_o + \beta_o A_o^{-1} - I)\} \leq \mathrm{tr}\{(-I)(-I)\} = n$ since α_o, β_o minimize $e(A_1)$.

(vii) This follows from the fact that $\beta(A_k) \geq 0$ and property (v).

(viii) Since A_k is normal, we have $\mathrm{Re}\{\mathrm{tr}(A_k^* A_k^{-1})\}$
$= \mathrm{Re}\{\sum_{i=1}^{n} (\bar{\lambda}_i(A_k)/\lambda_i(A_k))\} \geq n$.

(ix) By applying property (viii) to the first normal equation we have $2\mathrm{Re}\{\mathrm{tr}(A_k)\} \leq 2\alpha_k \mathrm{tr}(A_k^* A_k) + 2\beta_k n$. Since $\mathrm{tr}(A_k^* A_k) = \mathrm{Re}\{\mathrm{tr}(A_k)\} \leq n$, it follows that $\alpha_k + \beta_k \geq \{\mathrm{Re}(\mathrm{tr}(A_k))\}/n$.

Our next goal is to prove the following main convergence theorem.

THEOREM 4.3.2. Let

$$A_{k+1} = \alpha_k A_k + \beta_k A_k^{-1}, \quad k \in \mathbb{N}$$

with $A_o \in NS_n^+$ where α_k and β_k are given by (4.3.8) and (4.3.9). Then as $k \to \infty$, $A_k \to I$, $\alpha_k \to 1/2$ and $\beta_k \to 1/2$.

Proof. If, for certain k, A_k is a scalar matrix, the result is obvious because of Lemma 4.3.4. Thus we can assume that for every $k \in \mathbb{N}$, A_k is not a scalar matrix. We have therefore, $\alpha_k, \beta_k > 0$ for every $k \in \mathbb{N}$.

As A_o can be reduced to a diagonal matrix by unitary transformation, it is clear that there is no loss of generality in assuming that for every $k \in \mathbb{N}$, in the algorithm, A_k is diagonal. If we write $A_k = \mathrm{diag}(\lambda_1^{(k)}, \ldots, \lambda_n^{(k)})$, then $e_k = \sum_{i=1}^{n} |\lambda_i^{(k)} - 1|^2$. Hence the boundedness of $\{e_k\}$ implies that each one of the n sequences $\{\lambda_i^{(k)}\}$, $i=1,\ldots,n$ is bounded. Now by Cantor's diagonal process there exists a subsequence $\{k_p\}$ of nonnegative integers where $k_o < k_1 < k_2 < \ldots$ such that $\lambda_i^{(k_p)} \to \tilde{\lambda}_i$ say, as $p \to \infty$, for each $i=1,\ldots,n$. Since $\{\lambda_i^{(k_p)}\}$ is a sequence of complex numbers with positive real parts, $\mathrm{Re}(\tilde{\lambda}_i) \geq 0$, $i=1,\ldots,n$. Further, in the iterative process, due to the positive stability of the initial matrix A_o and

the positivity of α_k, β_k, a simple geometric reasoning reveals that $|\arg(\lambda_i^{(k+1)})| \leq |\arg(\lambda_i^{(k)})|$ for $i=1,\ldots,n$, where $\arg(\lambda)$ denotes the principal argument of λ so that $-\pi < \lambda \leq \pi$. In view of this, it follows that $\mathrm{Re}(\tilde{\lambda}_i) \geq 0$ and equality holds iff $\tilde{\lambda}_i = 0$. Consequently, we have any one and only of the following three cases:

(i) $\tilde{\lambda}_i = 0$ for all $i=1,\ldots,n$.

(ii) $\tilde{\lambda}_i = 0$ for at least one i and at most $(n-1)$ indices $i \in \{1,\ldots,n\}$ and $\mathrm{Re}(\tilde{\lambda}_j) > 0$ if $\tilde{\lambda}_j \neq 0$.

(iii) $\mathrm{Re}(\tilde{\lambda}_i) > 0$ for all $i=1,\ldots,n$.

If $\tilde{\lambda}_i = 0$ for all $i=1,\ldots,n$, by Lemma 4.3.4(c), we have $n > e_1 > \lim_{p \to \infty} e_{k_p} = n$ which is a contradiction.

Next, suppose $S=\{i_1,i_2,\ldots,i_m\}$, $1 \leq m \leq n-1$, to be the set of indices i for which $\tilde{\lambda}_i$ vanishes. Again by using Cantor's diagonal process we can always find a subsequence $\{k_{1,p}\}$ of $\{k_p\}$ such that for some $j \in \{1,\ldots,m\}$ all the m limits

$$\lim_{p \to \infty} \frac{\lambda_i^{(k_{1,p})}}{\lambda_j^{(k_{1,p})}}, \quad i \in S$$

exist and are bounded by unity in magnitude. We can always assume that $i_j=1$ and the remaining elements of S to be $2,3,\ldots,m$, since a permutation transformation does not affect the set up of the algorithm. Now let

$$\lim_{p \to \infty} \frac{\lambda_1^{(k_{1,p})}}{\lambda_i^{(k_{1,p})}} = c_i \text{ where } |c_i| \leq 1, \; i=1,\ldots,n. \tag{4.3.18}$$

In the following analysis, unless otherwise specified, the summations correspond to values of i from 1 to n. Putting $\lambda_i^{(k_1,p)} = g_{i,p}$, i=1,...,n it is easily seen that

$$\alpha_{k_1,p} = \frac{(\mathrm{Re}\Sigma g_{i,p})\Sigma|\frac{g_{1,p}}{g_{i,p}}|^2 - \{\mathrm{Re}\Sigma(g_{i,p}|\frac{g_{1,p}}{g_{i,p}}|^2)\}\{\mathrm{Re}\Sigma(\frac{\bar{g}_{i,p}}{g_{i,p}})\}}{\Sigma|g_{i,p}|^2\,\Sigma|\frac{g_{1,p}}{g_{i,p}}|^2 - |g_{1,p}|^2\{\mathrm{Re}\Sigma(\frac{\bar{g}_{i,p}}{g_{i,p}})\}^2} \tag{4.3.19}$$

We know that $g_{i,p} \to 0$ for i=1,...,m; $(g_{1,p}/g_{i,p}) \to c_i$ for i=1,...,m; and $(g_{1,p}/g_{i,p}) \to 0$ for i=m+1,...,n. Therefore we have

$$\lim_{p\to\infty} \alpha_{k_1,p} = \frac{\{1+\sum_{i=2}^{m}|c_i|^2\}\lim_{p\to\infty}\{\mathrm{Re}\sum_{i=m+1}^{n} g_{i,p}\}}{\{1+\sum_{i=2}^{m}|c_i|^2\}\lim_{p\to\infty}\{\sum_{i=m+1}^{n}|g_{i,p}|^2\}}$$

$$= \frac{\mathrm{Re}\sum_{i=m+1}^{n}\tilde{\lambda}_i}{\sum_{i=m+1}^{n}|\tilde{\lambda}_i|^2} = \ell, \text{ say.}$$

It may be noted that $\ell > 0$. By similar computations as above, it can be shown that

$$\lim_{p\to\infty} \beta_{k_1,p} = 0. \tag{4.3.20}$$

Hence for i=m+1,...,n,

$$\lim_{p\to\infty} \lambda_i^{(k_1,p+1)} = \lim_{p\to\infty}\{\alpha_{k_1,p}\lambda_i^{(k_1,p)} + \beta_{k_1,p}/\lambda_i^{(k_1,p)}\}$$

$$= \ell\tilde{\lambda}_i,$$

which has a positive real part.

Next we prove that the sequence $\{\lambda_1^{(k_{1,p}+1)}\}$ possesses a convergent subsequence whose limit has a positive real part. For this,

$$\lambda_1^{(k_{1,p}+1)} = \alpha_{k_{1,p}} \lambda_1^{(k_{1,p})} + \beta_{k_{1,p}} / \lambda_1^{(k_{1,p})} \qquad (4.3.21)$$

The first term on the right hand side of (4.3.21) tends to zero. For the second term, we easily verify that

$$\lim_{p \to \infty} \beta_{k_{1,p}} / \lambda_1^{(k_{1,p})}$$

$$= \lim_{p \to \infty} \frac{(1/g_{1,p}) \mathrm{Re}\Sigma(1/g_{i,p}) \Sigma |g_{i,p}|^2 - (1/g_{1,p})(\mathrm{Re}\Sigma g_{i,p}) \cdot \mathrm{Re}\Sigma(\bar{g}_{i,p}/g_{i,p})}{\Sigma |g_{i,p}|^2 \cdot \Sigma |g_{i,p}|^{-2} - \{\mathrm{Re}\Sigma(\bar{g}_{i,p}/g_{i,p})\}^2}$$

$$= \lim_{p \to \infty} \frac{\bar{g}_{1,p}\{\mathrm{Re}\Sigma(1/g_{i,p})\}\Sigma |g_{i,p}|^2 - \bar{g}_{1,p}\{\mathrm{Re}\Sigma g_{i,p}\} \cdot \mathrm{Re}\Sigma(\bar{g}_{i,p}/g_{i,p})}{\Sigma |g_{i,p}|^2 \Sigma |g_{1,p}/g_{i,p}|^2 - |g_{1,p}|^2 \{\mathrm{Re}\Sigma(\bar{g}_{i,p}/g_{i,p})\}^2} .$$

In the last expression the second terms of the numerator and the denominator tend to zero. The first term in the denominator converges to $\sum_{i=m+1}^{n} |\tilde{\lambda}_i|^2 (1 + \sum_{i=2}^{m} |c_i|^2)$. Since $\Sigma |g_{i,p}|^2 \to \sum_{i=m+1}^{n} |\tilde{\lambda}_i|^2 \neq 0$, to prove that the required limit exists and has a positive real part it is therefore sufficient to prove that

$$\lim_{p \to \infty} \bar{g}_{1,p} \, \mathrm{Re}(\Sigma(1/g_{i,p}))$$

exists and has a positive real part. If we now take $\mathrm{Re}(g_{i,p}) = a_{i,p}$ and $\mathrm{Im}(g_{i,p}) = b_{i,p}$ then

$$\lim_{p\to\infty} \mathrm{Re}\{\bar{g}_{1,p}\mathrm{Re}\sum_{i=1}^{n}(1/g_{i,p})\} = \lim_{p\to\infty} a_{1,p}\sum_{i=1}^{m}\frac{a_{i,p}}{a_{i,p}^2+b_{1,p}^2}.$$

Since $a_{i,p} > 0$ for $i=1,\ldots,n$,

$$a_{1,p}\sum_{i=1}^{m}\frac{a_{i,p}}{a_{i,p}^2+b_{i,p}^2} > \frac{a_{1,p}^2}{a_{1,p}^2+b_{1,p}^2}$$

$$= \frac{1}{1+\{\frac{b_{1,p}}{a_{1,p}}\}^2}$$

$$= \cos^2(\theta_{1,p})$$

where

$$\theta_{1,p} = |\tan^{-1}\{\frac{b_{1,p}}{a_{1,p}}\}|.$$

We know $\theta_{1,p} \le \theta_o < \pi/2$, where θ_o is the argument of the eigenvalue $\lambda_1(A_o)$. Hence

$$\mathrm{Re}\{\bar{g}_{1,p}\mathrm{Re}\sum_{i=1}^{m}(1/g_{i,p})\} > \cos^2(\theta_o) > 0 \qquad (4.3.22)$$

Now

$$|\bar{g}_{1,p}\sum_{i=1}^{m}\frac{a_{i,p}}{a_{i,p}^2+b_{i,p}^2}| \le \sum_{i=1}^{m}|g_{1,p}/g_{i,p}|$$

$$\to 1 + \sum_{i=2}^{m}|c_i| \quad \text{as } p\to\infty.$$

Hence the sequence $\{g_{1,p}\mathrm{Re}\sum_{i=1}^{n}(1/g_{i,p})\}$ is bounded and therefore possesses a convergent subsequence. In view of

(4.3.22) the real part of the limit of the subsequence is positive. Thus as has been the case with the proof of Theorem 4.2.1, in at most n-1 constructions of such subsequences, the case (ii) leads us to a consideration of the case (iii).

Thus assuming that $Re(\tilde{\lambda}_i) > 0$, i=1,...,n, we are to show that $A_k \to I$, or equivalently $e_k \to 0$. The proof of this closely resembles the analysis of the cases (iii) and (iv) of the proof of Theorem 4.2.1 and hence is omitted.

Finally, we are left to prove that $\alpha_k, \beta_k \to 1/2$. Since $\tilde{\lambda}_i \to 1$, i=1,...,n, for sufficiently large values k,

$$\lambda_i^{(k)} = 1 + \varepsilon_{i,k} \text{ where } \varepsilon_{i,k} \to 0 \text{ as } k \to \infty \text{ for } i=1,\ldots,n.$$

Let us use the notation

$$\begin{aligned} s_1 &= \Sigma\, \varepsilon_{i,k} \\ \bar{s}_1 &= \Sigma\, \bar{\varepsilon}_{i,k} \\ s_2 &= \Sigma\, \varepsilon_{i,k}^2 \\ \bar{s}_2 &= \Sigma\, \bar{\varepsilon}_{i,k}^2 \\ \tilde{s}_2 &= \Sigma\, \varepsilon_{i,k}\bar{\varepsilon}_{i,k} \end{aligned}$$

By straightforward calculations we have

$$\begin{aligned} \mathrm{tr}(A_k^* + A_k) &= 2n + s_1 + \bar{s}_1 \\ \mathrm{tr}(A_k^* A_k) &= n + s_1 + \bar{s}_1 + \tilde{s}_2 \\ \mathrm{tr}(A_k^{-*} + A_k^{-1}) &= 2n - s_1 - \bar{s}_1 + s_2 + \bar{s}_2 + \ldots. \\ \mathrm{tr}(A_k^{-*} A_k^{-1}) &= n - s_1 - \bar{s}_1 + s_2 + \bar{s}_2 + \tilde{s}_2 + \ldots \end{aligned}$$

and

$$\mathrm{tr}(A_k^* A_k^{-1} + A_k^{-*} A_k) = 2n + s_2 + \bar{s}_2 - 2\tilde{s}_2 + \ldots$$

where the neglected terms are of third and higher order. Therefore,

$$\tilde{\alpha}(A_k) = 8n\tilde{s}_2 - 2s_1^2 - 2\bar{s}_1^2 - 4s_1\bar{s}_1 + \cdots$$

$$\tilde{\beta}(A_k) = 8n\tilde{s}_2 - 2s_1^2 - 2\bar{s}_1^2 - 4s_1\bar{s}_1 + \cdots$$

and

$$\gamma(A_k) = 16n\tilde{s}_2 - 4s_1^2 - 4\bar{s}_1^2 - 8s_1\bar{s}_1 + \cdots$$

Now if we put $\mathrm{Re}(\varepsilon_{i,k}) = a_i$ and $\mathrm{Im}(\varepsilon_{i,k}) = b_i$ for convenience then

$$\begin{aligned} 8n\tilde{s}_2 - 2s_1^2 - 2\bar{s}_1^2 - 4s_1\bar{s}_1 &= 8n\,\Sigma(a_i^2+b_i^2) - 8(\Sigma a_i)^2 \\ &= 8\{n\Sigma a_i^2 - (\Sigma a_i)^2\} + 8n\Sigma b_i^2 \\ &\geq 0 \end{aligned}$$

with equality holding iff $b_i=0$ for all $i=1,\ldots,n$ and $a_i = a_j$ for all pairs (i,j). Since A_k is not a scalar matrix this will not happen. Hence $8n\tilde{s}_2 - 2s_1^2 - 2\bar{s}_1^2 - 4s_1\bar{s}_1 > 0$.

Thus $\alpha(A_k) = \tilde{\alpha}(A_k)/\gamma(A_k) \to 1/2$. Similarly $\beta(A_k) \to 1/2$. This completes the proof of the main convergence theorem. ▲▲▲

REMARK 4.3.1. The result of the above theorem remains valid even if the assumption $A \in NS_n^+$ is replaced by $A \in NS_n^-$ where $NS_n^- = N_n \cap S_n^-$. For, as may be easily verified, in this case $A_1 \in NS_n^+$.

REMARK 4.3.2. If A is Hermitian the algorithms of this section and the previous section are one and the same.

REMARK 4.3.3. In contradistinction to Remark 4.2.2, in the case of the present algorithm the convergence for the case of a 2x2 matrix need not be in just one step. This is

easily seen by taking $A = \begin{bmatrix} 1+i & 0 \\ 0 & 1 \end{bmatrix}$. By direct computations $\alpha_0 = 3/7$ and $\beta_0 = 5/7$ and therefore,

$$A_1 = \begin{bmatrix} \frac{(11+i)}{14} & 0 \\ 0 & \frac{8}{7} \end{bmatrix}.$$

Our numerical simulations confirm that the algorithm presented in this section is considerably faster. We find that the algorithm requires three to five iterations. For the following two matrices $A \in NS_n^+$ we verified our method and in each case it took only three iterations to achieve an accuracy of six decimal places:

$$\begin{bmatrix} 7 & 3 & -1 & 2 \\ 2 & 7 & 3 & -1 \\ -1 & 2 & 7 & 3 \\ 3 & -1 & 2 & 7 \end{bmatrix} \qquad \begin{bmatrix} 10 & -1 & 5 & 1 & 1 & 0 \\ 0 & 10 & -1 & 5 & 1 & 1 \\ 1 & 0 & 10 & -1 & 5 & 1 \\ 1 & 1 & 0 & 10 & -1 & 5 \\ 5 & 1 & 1 & 0 & 10 & -1 \\ -1 & 5 & 1 & 1 & 0 & 10 \end{bmatrix}.$$

For certain nonnormal matrices also we just tried the method. In certain cases the method converges. For example, by taking

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 7 & 4 \\ 0 & 0 & -2 & 3 \end{bmatrix}$$

we find that the method converges in four iterations.

4.4. The Kaczmarz Projection Method for Solving AX+XB=C

The well-established Bartels-Stewart algorithm for solving

$$AX + XB = C \tag{4.4.1}$$

is a direct method, while in general, iterative methods are preferred for solving large systems. The rapidly convergent iterative method suggested by Hoskins, Meek and Walton [205] as described earlier involves inversions of two matrices in each iteration and this method is applicable only when either A and B or -A and -B are stable. Moreover, it is well-known that the inversion of very large-order or nearly singular matrices is rather unwieldy. Furthermore, in the above method the original matrices A and B are altered as the computations proceed and there is no particular advantage to be gained when A and B are sparse, i.e., have a large number of zero elements or that the matrices have some systematic arrangement in their entries so that they can be easily generated without having to be stored in the core. Systems involving large-order sparse matrices arise, for example from the finite-difference approximations for partial differential equations as we have seen in the introductory chapter. Another iterative method

$$AX_{k+1} = -X_k B + C \tag{4.4.2}$$

for solving (4.4.1) mentioned in Varah [272] converges iff $\rho(B^T \otimes A^{-1}) < 1$. In any case, efficient algorithms have seldom been reported for singular or inconsistent cases [305].

Keeping the above points in mind, in this and the following section we propose two iterative schemes to solve (4.4.1) for large sparse systems even when the system is singular or inconsistent. These schemes are based on our observation that the projection (due to Kaczmarz [296]) and the residual projection methods (due to Nekrasov [299] (refer Forsythe [290] and Hestenes and Stein [294])) could be very efficiently implemented forthe matrix equation AX+XB=C, as well as its other generalizations mentioned in Chapter 1. For definiteness, however, we shall restrict ourselves to the case of the Sylvester equation only. The theoretical unconditional convergence and the characterization of the solution so obtained for these methods for the case of a general linear system $Ax = b$ ($A \in M_{m,n}$, $b \in \mathbb{C}^m$) have been established respectively in Tanabe [305] and Rathore [301].

In either of these methods we generate a sequence of matrices $\{X_k\}$ which converges for arbitrary A, B, C and X_0. In both the methods, the limit of the sequence, say X_∞ represents the unique solution in case the system is nonsingular (i.e. if A and -B do not have a common eigenvalue). If the system is singular but consistent, then in both methods, X_∞ represents some solution of the system. In particular if we choose $X_0 = 0$ then the Kaczmarz projection method gives the minimum Frobenius norm solution. On the other hand,

if the system is inconsistent then X_∞ obtained by the residual projection method corresponds to a least squares solution (i.e., $X = X_\infty$ minimizes $\|C-AX-XB\|$). In this context, it may be referred that Lovass-Nagy and Powers [233] have suggested a method to compute least squares solutions and the method relies on a strategy of replacing C by G which is the orthogonal projection of C on the range of the operator A.+.B, so that the new equation AX+XB=G becomes consistent and then solving this consistent system by some direct method.

Some advantages and disadvantages of the Kaczmarz and the residual projection methods are as follows. In both these methods the convergence of the process is theoretically guaranteed. However as is the case with several first order iterative methods, in certain cases the rate of convergence could be rather slow. Thus, as a rule, the methods are recommended only for large systems, where these have a better chance of comparing favourably with direct methods. A practical strategy would be to first test run the method to see if the convergence is quick. If not, one has to introduce suitable modifications such as relaxation factors after observing the iterates in a known solution case, to accelerate the convergence.

Now we shall formulate the Kaczmarz projection method for solving (4.4.1). As a prelude, we first describe the basic projection algorithm [305] for solving the linear system

$Ax = b$, where $A = (a_{ij}) \in M_{m,n}$, $x \in \mathbb{C}^n$ and $b = (b_1,\ldots,b_m)^T \in \mathbb{C}^m$. Here it is assumed that all the rows of A are nonzero. As at earlier occasions we use $A^{(j)}$ to denote the j-th column of A and $A_{(i)}$ to denote the i-th row of A. Throughout this section and the next one, $A^*_{(j)}$ and $B^{(j)*}$ denote the conjugate transpose of $A_{(i)}$ and $B^{(j)}$ respectively. In the case of vectors, x^k denotes the k-th iterate of x and in the case of matrices, X_k denotes the k-th iterate of X. Let

$$f_i(x) = x - \frac{(x, A^*_{(i)}) - b_i}{\alpha_i} A^*_{(i)}, \quad i=1,..,m \qquad (4.4.3)$$

$$F(b,x) = f_1(f_2(f_3(..(f_m(x))...))) \qquad (4.4.4)$$

where

$$\alpha_i = (A^*_{(i)}, A^*_{(i)}), \quad i=1,\ldots,m, \qquad (4.4.5)$$

(u,v) denoting the standard inner product v^*u. It may be observed that

$$f_i(x) = P_i x + \frac{b_i}{\alpha_i} A^*_{(i)}, \quad i=1,\ldots,m \qquad (4.4.6)$$

where

$$P_i = I - \frac{1}{\alpha_i} A^*_{(i)} A_{(i)}, \quad i=1,\ldots,m, \qquad (4.4.7)$$

I being the n x n identity matrix. Let us define

$$Q_i = P_1 P_2 \ldots P_i, \; Q_o = I, \; i=1,\ldots,m \qquad (4.4.8)$$

and let us denote Q_m by Q. Let P_I and P_K respectively denote the orthogonal projections onto the subspaces Im A^* and Ker A, where the symbols Im and Ker stand for the image (range) and Kernel (null space) of the corresponding mappings. Further,

define G as the $n \times m$ matrix $(I-QP_I)^{-1}R$ where R is the $n \times m$ matrix whose i-th column is $Q_{i-1}A^*_{(i)}/\alpha_i$, i=1,...,m. (The theory developed in Tanabe [305] guarantees that $I-QP_I$ is invertible.) Choose an arbitrary vector $x^0 \in \mathbb{C}^n$ and determine the sequence $\{x^k\}$ from the recurrence relation

$$x^{k+1} = F(b,x^k), \quad k=0,1,2,... \tag{4.4.9}$$

Then it has been established by Tanabe [305] through a chain of results that

$$\lim_{k \to \infty} x^k = P_K x^0 + Gb. \tag{4.4.10}$$

If the system Ax = b is consistent then the above limit is a solution of the system for arbitrary x^0. If we choose $x^0 = 0$ then the above iterative method provides the minimum norm solution. Even if the system is inconsistent, the above limit exists. However, that need not be a least squares solution.

The total number of multiplicative operations involved in the above method is $\{(2s+1)mk+s\}$ where k is the number of iterations and s is the number of nonzero elements in A. The above algorithm has a simple geometric interpretation. By the mapping f_i, a vector $x \in \mathbb{C}^n$ is projected on the hyperplane defined by $A_{(i)}x = b_i$. Then F(b,x) is obtained from x, being projected succesively on the hyperplanes $A_{(i)}x = b_i$, in the order i=m,m-1,...,1. This cycle forms a single iteration of the algorithm. In fact, if the hyperplanes are mutually orthogonal, then a single iteration yields an answer.

Based on the above method let us now describe a procedure to solve AX+XB=C with $A = (a_{ij}) \in M_m$, $B = (b_{ij}) \in M_n$, $C = (c_{ij}) \in M_{m,n}$ and $X = (x_{ij}) \in M_{m,n}$. Of course, we have already seen that the equation (4.4.1) can be recast as a linear system

$$\begin{bmatrix} A+b_{11}I_m & b_{21}I_m & \cdots & b_{n1}I_m \\ b_{12}I_m & A+b_{22}I_m & \cdots & b_{n2}I_m \\ \vdots & \vdots & & \vdots \\ b_{1n}I_m & b_{2n}I_m & \cdots & A+b_{nn}I_m \end{bmatrix} \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(n)} \end{bmatrix} = \begin{bmatrix} C^{(1)} \\ C^{(2)} \\ \vdots \\ C^{(n)} \end{bmatrix} \qquad (4.4.11)$$

where I_m is the m x m identity matrix, $X^{(j)}$ and $C^{(j)}$ $(j=1,\ldots,n)$ are the j-th columns of X and C respectively. The matrix in the left hand side of (4.4.11) is called the Kronecker sum of A and B defined by

$$M = (I_n \otimes A) + (B^T \otimes I_m). \qquad (4.4.12)$$

Now we shall apply the Kaczmarz algorithm to solve the enlarged system and then revert back to the compact form. In this procedure we slightly deviate from the ordering used in Tanabe's paper in the sense that we consider the hyperplanes in the order corresponding to $c_{11}, c_{21}, \ldots, c_{m1}, c_{12}, \ldots, c_{m2}, \ldots, c_{1n}, \ldots, c_{mn}$ and in Tanabe's analysis it is just in the reverse order. There is no loss of generality in the above assumption as now only the first equation is considered first and so on. To formulate the algorithm, the only thing required is to identify the hyperplane corresponding to c_{ij} in the enlarged system. The basic step is to write the analogue of (4.4.3)

For this, let us define

$$\alpha_{ij} = (A^{*}_{(i)}, A^{*}_{(i)}) + (B^{(j)}, B^{(j)}) + 2\,\mathrm{Re}(\bar{a}_{ii} b_{jj}) \qquad (4.4.13)$$

and

$$\beta_{ij} = (X^{(j)}, A^{*}_{(i)}) + (B^{(j)}, X^{*}_{(i)}) - c_{ij}. \qquad (4.4.14)$$

To facilitate the presentation, we shall also define the following two mappings. Let

$$T^{(j)}: M_{m,1} \to M_{m,n}, \quad j=1,\ldots,n$$

such that

$$T^{(j)}(u) = U$$

where

$$U^{(j)} = u \text{ and } U^{(i)} = 0, \; i \neq j.$$

In other words $T^{(j)}(u)$ is an $m \times n$ matrix whose j-th column is u and all other columns are zero. Similarly let

$$T_{(i)}: M_{1,n} \to M_{m,n}, \quad i=1,\ldots,m$$

such that

$$T_{(i)}(v) = V$$

where

$$V_{(i)} = v \text{ and } V_{(j)} = 0, \; j \neq i.$$

Therefore, $T_{(i)}(v)$ is an $m \times n$ matrix whose i-th row is v and all other rows are zero. It can be easily seen that the analogue of (4.4.3) is

$$f_{ij}(X) = X - d_{ij}\{T^{(j)}(A^{*}_{(i)}) + T_{(i)}(B^{(j)*})\} \qquad (4.4.15)$$

where

$$d_{ij} = \beta_{ij}/\alpha_{ij}, \text{ provided } \alpha_{ij} \neq 0. \qquad (4.4.16)$$

If $\alpha_{ij}=0$, let us take

$$f_{ij}(X) = X. \tag{4.4.17}$$

It may be noted that if $\alpha_{ij}=0$, then A and -B have a common eigenvalue. Thus vanishing of any α_{ij} is an indication for the system (4.4.1) to have more than one solution or no solution. However, vanishing of any α_{ij} is not a necessary condition for the system to be singular. Let

$$F(C,X) = f_{mn}(f_{m-1,n}(\ldots(f_{1n}(\ldots(f_{m1}(f_{m-1,1}\ldots(f_{11}(X))\ldots))\ldots)) \ldots)). \tag{4.4.18}$$

Now choosing an initial approximation $X_o \in M_{m,n}$, let us generate the sequence of matrices $\{X_k\}$ by the relation

$$X_{k+1} = F(C,X_k), \quad k=0,1,2,\ldots \tag{4.4.19}$$

Since this sequence can be identified with the sequence of vectors in $\mathcal{T}^{mn}$, as an immediate consequence of the convergence result established by Tanabe [305], it now follows that $\lim_{k\to\infty} X_k$ exists and it represents a solution of (4.4.1) in case the system is consistent. If we consider the initial approximation as $X_o = 0$, then the above limit gives the minimum norm solution.

Now we shall present the above algorithm suited to computer implementation. Here the algorithm is considered for the real case.

Step 1. Choose any X_o, for example $X_o=0$, and set $X=X_o$.

Step 2. Compute for $i=1,\ldots,m$

$$p_i = \sum_{s=1}^{m} a_{is}^2 .$$

Step 3. Compute for $j=1,\ldots,n$

$$q_j = \sum_{t=1}^{n} b_{tj}^2.$$

Step 4. Set $k=1$.

Step 5. For $j=1,\ldots,n$, $i=1,\ldots,m$ do:

Skip this step for those j and i when

$$p_i+q_j+2a_{ii}b_{jj} = 0.$$

$$\text{Set } \beta = \sum_{s=1}^{m} a_{is}x_{sj} + \sum_{t=1}^{n} x_{it}b_{tj} - c_{ij}$$

$$d = \beta/(p_i+q_j+2a_{ii}b_{jj})$$

For $s=1,\ldots,m$ update x_{sj} as $x_{sj}-da_{is}$

For $t=1,\ldots,n$ update x_{it} as $x_{it}-db_{tj}$.

Step 6. If the matrix X (which is now the k-th iterate X_k) satisfies an acceptance criterion for a convergence, go to Step 7.

Step 7. The process is complete.

In the above algorithm, as in the Gauss-Seidel iteration, we renovate X successively in each iterative step. Hence it requires only one iterate to be kept in memory. However, if we consider the stopping criterion for Step 6 as to iterate until

$$\|X_k - X_{k-1}\| < \varepsilon \qquad (4.4.20)$$

or

$$\frac{\|X_k - X_{k-1}\|}{\|X_k\|} < \varepsilon \qquad (4.4.21)$$

$\varepsilon > 0$ being some prescribed tolerance, then we require mn memory locations to keep the previous iterate. Another stopping criterion involving two successive iterates is to see whether the norm of the residual, i.e. $\|C-AX_k-X_kB\|$ is very close to that corresponding to the previous iterate. In case the system has a solution, then

$$\|C-AX_k-X_kB\| < \varepsilon \qquad (4.4.22)$$

can be considered to terminate the iterative process. These criteria involving residuals require extra multiplications compared with (4.4.20). It may be noted that $\|X\|^2 = \sum_{i,j} |x_{ij}|^2$. We have followed the stopping criterion as (4.4.20) in our numerical examples which will be presented in the end of this section.

We shall now discuss some computational aspects of the above algorithm. The algorithm is very simple to implement on a computer. The algorithm does not involve inversion of matrices or transformation of matrices to canonical forms. Calculation of p_i and q_j values involves m^2+n^2 multiplications. In each iteration, there are at most mn steps and each step requires 2m+2n+3 multiplications (m+n+2 for β, 1 for d and m+n for updating X). Hence each iteration requires (2m+2n+3)mn multiplications whereas the Hoskins-Meek-walton algorithm [205] requires $\frac{4}{3}(m^3+n^3) + mn(m+n)$ multiplications per iteration. In the present algorithm, calculations in stopping criterion involve mn multiplications. Therefore, the total number

of multiplicative operations involved in the process is at most $m^2+n^2+2mnk(m+n+2)$, k being the number of iterations required. For sparse matrices, this figure is reduced considerably.

It can be easily seen that the storage requirement for the proposed algorithm is $m^2+n^2+2mn+m+n$. It also requires additional mn locations for the previous iterate involved in the stopping criterion. If we do not count these mn locations then definitely for large systems this method requires smaller number of memory locations compared to the Bartels-Stewart algorithm and the Hoskins-Meek-Walton algorithm which require $2m^2+2n^2+mn$ and $(m+n)^2+\max(m^2,n^2)$ locations respectively. Even if we take the storage requirement for the present algorithm as $m^2+n^2+3mn+m+n$, we are in a favourable situation in most cases (that is if m/n is different from 1 and at least one of m and n is sufficiently large), for example m=2n or n=2m.

In conclusion, the Kaczmarz projection method proposed in this section is applicable to all classes of A and B. If $\|C-AX_\infty-X_\infty B\|$ is not close to zero then immediately we can infer that the system is inconsistent. Moreover this method seems less vulnerable to the growth of round-off errors. On the other hand, the convergence of the method is generally very slow compared to the Hoskins-Meek-Walton algorithm. Our computational experiments show that in some cases the total number of iterations required for convergence even to three or four decimal places is several hundred with each A and B

of size 10 x 10. However, there are systems which take only a small number of iterations. Some such examples follow. In these examples we give A, B and C and also the actual solution. To conserve space, the iterates X_k are given only for selected values k. In all these examples $X_o = 0$ and the stopping criterion is taken as $\|X_k - X_{k-1}\| < \varepsilon$ with $\varepsilon = 10^{-3}$.

EXAMPLE 4.4.1. $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $B = \begin{bmatrix} -2 & -1 \\ 2 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 3 & 3 \\ 2 & 0 \end{bmatrix}$

The system AX+XB=C is singular and consistent and therefore has infinitely many solutions.

$$X_1 = \begin{bmatrix} -0.599999 & 1.200000 \\ 0.300000 & 0.100000 \end{bmatrix}$$

$$X_{56} = \begin{bmatrix} -0.599999 & 1.200000 \\ 2.991783 & 0.997261 \end{bmatrix} \quad \text{and} \quad \|X_{56}\|^2 = 11.7453$$

From the output it is observed that the method converges to the solution $\begin{bmatrix} -0.6 & 1.2 \\ 3.0 & 1.0 \end{bmatrix}$.

EXAMPLE 4.4.2. We consider A and B as in the above example but with $C = \begin{bmatrix} 3 & -2 \\ 2 & 0 \end{bmatrix}$ so that the system AX+XB=C is inconsistent.

We find that X_k is approaching $\begin{bmatrix} 0.4 & -0.8 \\ 3.0 & 1.0 \end{bmatrix}$.

When k=41, $X_k = \begin{bmatrix} 0.400000 & -0.800000 \\ 2.960092 & 0.986697 \end{bmatrix}$

Also $\|X_{41}\|^2 = 10.5357$.

The above examples show that the convergence is very slow even for 2 x 2 system. As pointed out earlier, iterative methods are usually applied to large linear systems with a sparse coefficient matrix. Now we shall present some examples involving 10 x 10 matrices.

EXAMPLE 4.4.3.

$$A=\begin{bmatrix} 1 & 3 & 0 & 6 & 7 & 0 & 6 & 0 & 8 & 5 \\ 9 & 7 & 2 & 6 & 3 & 3 & 0 & 2 & 0 & 7 \\ 6 & 0 & 7 & 8 & 1 & 0 & 5 & 1 & 4 & 1 \\ 2 & 1 & 4 & 5 & 6 & 8 & 1 & 7 & 8 & 0 \\ 1 & 2 & 8 & 2 & 5 & 5 & 7 & 1 & 9 & 9 \\ 2 & 3 & 3 & 4 & 6 & 8 & 9 & 4 & 2 & 0 \\ 5 & 5 & 1 & 1 & 7 & 2 & 4 & 2 & 8 & 6 \\ 8 & 9 & 8 & 0 & 1 & 2 & 8 & 2 & 1 & 2 \\ 0 & 0 & 3 & 2 & 0 & 5 & 8 & 8 & 1 & 3 \\ 3 & 7 & 4 & 3 & 1 & 3 & 4 & 5 & 1 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} 2 & 0 & 7 & 2 & 5 & 3 & 4 & 1 & 9 & 5 \\ 6 & 2 & 4 & 9 & 5 & 1 & 0 & 3 & 5 & 5 \\ 5 & 1 & 8 & 1 & 7 & 4 & 0 & 1 & 2 & 0 \\ 7 & 2 & 5 & 3 & 4 & 6 & 0 & 2 & 2 & 6 \\ 1 & 4 & 6 & 5 & 8 & 0 & 0 & 2 & 9 & 6 \\ 2 & 9 & 6 & 3 & 8 & 4 & 0 & 3 & 1 & 1 \\ 6 & 9 & 2 & 5 & 1 & 4 & 0 & 2 & 2 & 5 \\ 0 & 1 & 8 & 0 & 5 & 2 & 2 & 2 & 1 & 7 \\ 5 & 3 & 7 & 1 & 1 & 7 & 9 & 7 & 6 & 2 \\ 5 & 5 & 5 & 5 & 0 & 3 & 1 & 5 & 5 & 9 \end{bmatrix}$$

C is chosen as AJ+JB where J is the matrix of unities, i.e., the matrix whose elements are all 1.

After 36 iterations, the termination criterion is satisfied, i.e. $\|X_{36} - X_{35}\| < 10^{-3}$. We find that $\|X_{36}\|^2 = 100.00396$. The matrix given below is X_k when k=36. (We give only upto three decimal places)

$$\begin{bmatrix} 1.000 & 1.010 & 0.996 & 1.007 & 1.004 & 1.004 & 0.999 & 1.001 & 0.991 & 0.991 \\ 0.998 & 0.999 & 1.005 & 0.994 & 1.000 & 0.996 & 0.998 & 1.000 & 1.003 & 1.007 \\ 1.003 & 1.000 & 1.003 & 1.002 & 1.012 & 0.998 & 0.993 & 0.990 & 0.993 & 1.005 \\ 0.999 & 0.999 & 1.003 & 0.990 & 0.991 & 1.006 & 1.012 & 1.005 & 0.997 & 1.001 \\ 0.996 & 0.995 & 0.992 & 1.011 & 1.009 & 1.002 & 0.996 & 0.989 & 1.002 & 1.000 \\ 0.995 & 0.989 & 1.003 & 1.002 & 1.005 & 0.997 & 1.002 & 1.000 & 1.003 & 1.002 \\ 1.002 & 1.006 & 0.999 & 1.002 & 0.996 & 1.001 & 1.001 & 1.005 & 0.997 & 0.997 \\ 1.004 & 0.998 & 0.992 & 1.005 & 1.000 & 0.996 & 0.993 & 0.996 & 1.010 & 1.000 \\ 1.002 & 1.002 & 1.008 & 0.988 & 0.990 & 1.001 & 1.009 & 1.011 & 1.000 & 0.998 \\ 0.998 & 1.000 & 1.000 & 1.000 & 0.997 & 1.001 & 1.000 & 0.997 & 1.003 & 1.000 \end{bmatrix}$$

EXAMPLE 4.4.4. Let $A = (a_{ij})$ be defined by

$$a_{ij} = \begin{cases} i & \text{if } i=j, \\ 1 & \text{if } |i-j|=1, \\ 0 & \text{otherwise.} \end{cases}$$

By considering B as A and C as AJ+JB where J is the matrix of unities we find that the iteration terminates in 21 steps. $\|X_{21}\|^2 = 99.8269$ and $\|X_{21}-X_{20}\| = 1.3 \times 10^{-6}$.

Finally let us consider a system AX+XB=C whose enlarged system is tridiagonal.

EXAMPLE 4.4.5. Let A be defined as in the preceding example. Let B = diag(1,2,...,10) and C=AJ+JB where J is the matrix of unities. In this case we find that $\|X_8\|^2 = 99.998758$ and all the elements in X_8, except the first three elements in the first column, are either 0.99999 or 1.00000. The first three elements in the first column are respectively 0.99879, 1.00082 and 1.00007.

In the light of above examples, we recommend that the Kaczmarz projection method as explained in this section may be used to solve large sparse systems.

4.5. The Residual Projection Method for Solving AX+XB=C

In this section, we formulate the residual projection method for solving AX+XB=C with A, B, C, X as assumed in the preceding section. The key to this method is once again to consider the enlarged system (4.4.11) and then to apply the residual projection method analysed in Rathore [301] for

solving linear systems and then to rewrite the computations in the compact form suited to the original matrix equation.

To begin with let us briefly outline with some simplifications, the residual projection method as given in Rathore [301] for solving $Ax = b$, $A \in M_{m,n}$. (The method as considered in [301] incorporates a general inner product on $\mathbb{C}^m$ whereas in the following it has been specialised to the usual one, i.e., $(u,v) = v^*u$.)

The residual projection algorithm generates a sequence of vectors $\{x^k\}$ in which the j-th component x_j^k of x^k is computed through

$$d_j^{k+1} = (r^{k,j-1}, A^{(j)})/\alpha_j \tag{4.5.1}$$

$$\alpha_j = (A^{(j)}, A^{(j)}) \tag{4.5.2}$$

$$x_j^{k+1} = x_j^k + d_j^{k+1} \tag{4.5.3}$$

and

$$r^{k,j} = r^{k,j-1} - d_j^{k+1} A^{(j)} \quad j=1,\ldots,n \tag{4.5.4}$$

where

$$r^{k+1} = r^{k+1,0} = r^{k,n}, \quad k=0,1,2\ldots \tag{4.5.5}$$

and

$$r^{0,0} = r^0 = b - Ax^0 \tag{4.5.6}$$

It has been established in [301] that the vector sequences $\{r^k\}$ and $\{x^k\}$ thus generated converge for arbitrary A, b and x^0 with

$$r^\infty = \lim_{k \to \infty} r^k = P_K b \tag{4.5.7}$$

and

$$x^\infty = \lim_{k \to \infty} x^k = Px^0 + Gb \tag{4.5.8}$$

where P_K is the orthogonal projection onto Ker A*, P is the projection onto Ker A along Im T, with T being the conjugate transpose of the m x n matrix whose j-th column is $S_{j-1}A^{(j)}/\alpha_j$, j=1,...,n. Here

$$S_j = R_1^*R_2^*\ldots R_j^*\ , \ S_o = I_m, \ S = S_n^* \tag{4.5.9}$$

$$R_j = I_m - A^{(j)}A^{(j)*}/\alpha_j. \tag{4.5.10}$$

In (4.5.8), $G = T(I_m - SP_I)^{-1}$ with P_I denoting the orthogonal projection onto Im A. Here the invertibility of $I_m - SP_I$ is guaranteed through the theory developed in [301]. It has also been characterized in [301] that x^∞ given by (4.5.8) minimizes $\|b-Ax\|$.

Now we shall apply the above method to the enlarged system (4.4.11) and rewrite the calculations suited to the original form (4.4.1). The algorithm can be described as follows. Let

$$\alpha_{ij} = (A^{(i)},A^{(i)}) + (B_{(j)}^*,B_{(j)}^*) + 2\,\mathrm{Re}(\bar{a}_{ii}b_{jj}) \tag{4.5.11}$$

$$\beta_{ij} = (R^{(j)},A^{(i)}) + (R_{(i)}^*,B_{(j)}^*) \tag{4.5.12}$$

where R is the updated residual C-AX-XB. Let

$$d_{ij} = \begin{cases} \beta_{ij}/\alpha_{ij} & \text{if } \alpha_{ij} \neq 0, \\ 0 & \text{if } \alpha_{ij} = 0. \end{cases} \tag{4.5.13}$$

Then update x_{ij} by adding d_{ij} to the existing x_{ij} and update the residual by $R - d_{ij}\{T^{(j)}(A^{(i)}) + T_{(i)}(B_{(j)})\}$, R being the existing residual and $T^{(j)}$ and $T_{(i)}$ are the mappings defined as in the previous section. If we complete these steps for

$j=1,\ldots,n$, $i=1,\ldots,m$ then we say one iteration is over. We continue the process until the required accuracy is achieved.

The step-by-step procedure of the above algorithm is presented below for the real case.

Step 1. Choose any X_o, for example $X_o=0$ and set $X=X_o$.

Step 2. Set $R = (r_{ij}) = C-AX-XB$

Step 3. Compute for $i=1,\ldots,m$

$$p_i = \sum_{s=1}^{m} a_{si}^2.$$

Step 4. Compute for $j=1,\ldots,n$

$$q_j = \sum_{t=1}^{n} b_{jt}^2.$$

Step 5. Set $k=1$.

Step 6. For $j=1,\ldots,n$, $i=1,\ldots,m$ do:

Skip this step for those j and i when

$$p_i+q_j+2a_{ii}b_{jj} = 0.$$

$$\text{Set } \beta = \sum_{s=1}^{m} a_{si}r_{sj} + \sum_{t=1}^{n} b_{jt}r_{it}$$

$$d = \beta/(p_i+q_j+2a_{ii}b_{jj})$$

Update x_{ij} as $x_{ij}+d$.

For $s=1,\ldots,m$ update r_{sj} as $r_{sj}-da_{si}$.

For $t=1,\ldots,n$ update r_{it} as $r_{it}-db_{jt}$.

Step 7. If the matrix X (which is now the k-th iterate X_k) or the matrix R (which is now the k-th iterate R_k) satisfies an acceptance criterion for a convergence go to Step 8.

EXAMPLE 4.5.1. Consider A, B, C as in Example 4.4.1. By the residual projection method we find that $\{X_k\}$ converges to $\begin{bmatrix} -3 & 0 \\ 3 & 1 \end{bmatrix}$ in 22 iterations. It may be observed that this is not a minimum norm solution.

EXAMPLE 4.5.2. Consider A, B, C as in Example 4.4.2. We find that $\|R_k\|$ converges to 12.5 and in this case $\{X_k\}$ converges to $\begin{bmatrix} -0.5 & 0.0 \\ 3.0 & 1.0 \end{bmatrix}$. X_{36} is given as

$$\begin{bmatrix} -0.5000000 & 0.0000000 \\ 2.9999923 & 0.9999982 \end{bmatrix}$$

with $\|X_{36}\|^2 = 10.249951$.

For the matrices A, B and C as given in Example 4.4.3 we executed the residual projection algorithm. We noted that the number of iterations exceeded eight thousands whereas by the Kaczmarz projection method, the solution was obtained in 36 iterations.

EXAMPLE 4.5.3. Consider A, B, C as in Example 4.4.4. After 91 iterations, the termination criterion is satisfied, i.e., $\|C-AX_{91}-X_{91}B\|^2 < 10^{-6}$. X_{91} is given below upto three decimal places.

$$\begin{bmatrix}
1.001 & 0.999 & 1.000 & 1.000 & 0.996 & 1.020 & 0.869 & 1.015 & 0.998 & 1.000 \\
0.999 & 1.001 & 1.000 & 1.000 & 1.001 & 0.995 & 1.015 & 0.997 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.001 & 0.998 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000
\end{bmatrix}$$

EXAMPLE 4.5.4. For the matrices A, B, C as given in Example 4.4.5, we find by the residual projection method that $\|R_k\|$ becomes less than 10^{-3} in 15 iterations.

REFERENCES

Books:

[1] S.Barnett, Matrices in control theory with applications to linear programming, Van Nostrand Reinhold Co., London (1971), 221 pp. Zbl. 245.93002

[2] S.Barnett, Introduction to mathematical control theory, Clarendon press, Oxford (1975), 264 pp. MR 55 ≠≠ 14276, Zbl. 307.93001.

[3] S.Barnett and C.Storey, Matrix methods in stability theory, Nelson, London (1970), 148 pp. Zbl.243.93017.

[4] R.Bellman, Introduction to matrix analysis, 2nd edn., McGraw-Hill, New York (1970), 403 pp. MR 41 ≠≠ 3493, Zbl. 216.61

[5] E.O.Brigham, The fast Fourier transform, Prentice-Hall, Englewood Cliffs, N.J. (1974), 252 pp.

[6] J.N.Franklin, Matrix theory, Prentice-Hall, Englewood Cliffs, N.J. (1968), 292 pp, MR 38 ≠≠ 5798, Zbl. 174.315.

[7] F.R.Gantmacher, The theory of matrices, Vols. I and II (trans. K.A.Hirsch), Chelsea, New York (1959), 374 pp. and 276 pp. MR 21 ≠≠ 6372 c.

[8] F.R.Gantmacher, Applications of the theory of matrices, Wiley Interscience, New York (1959), 317 pp. MR 21 ≠≠ 6372 b, Zbl.85.10.

[9] F.R.Gantmacher and M.G.Krein, Oscillatory matrices and Kernels and small vibrations of mechanical systems, 2nd edn., State publishing house for technical-theoretical literature, Moscow (1950), 408 pp. MR 22 ≠≠ 5161.

[10] A.S.Householder, The theory of matrices in numerical analysis, Blaisdell, New York (1964), 257 pp. MR 30 ≠≠ 5475, Zbl. 161.121.

[11] S.Karlin, Total positivity, Vol. I, Stanford University Press, Stanford, California (1968), 576 pp.

[12] M.G.Krein, Lectures on stability theory in the solution of differential equations in a Banach space, Inst. of Math., Ukrainian Acad. Sci., Kiev (1964) (in Russian).

[13] P.Lancaster, Theory of matrices, Academic Press, New York (1969), 316 pp. Zbl. 186.53.

[14] S.H.Lehnigk, Stability theorems for linear motions with an introduction to Liapunov's direct method, Prentice-Hall, Englewood,Cliffs, N.J. (1966), 251 pp.

[15] C.C.MacDuffee, The theory of matrices, Chelsea, New York (1956), 110 pp.

[16] M.Marcus and H.Minc, A survey of matrix theory and matrix inequalities, Allyn and Bacon, Boston (1964), 180 pp. MR 29 ≠≠ 112, Zbl. 126.24.

[17] M.Marden, Geometry of polynomials, Mathematical surveys, No.3, American Mathematical Society, Providence, R.I. New York (1966), 243 pp. MR 37 ≠≠ 1562, Zbl. 162.371.

[18] L.Mirsky, An introduction to linear algebra, Clarendon Press, Oxford (1961), 440 pp.

[19] T.Takahashi, Mathematics of automatic control (trans. Ed. G.M.Kranc), Holt, Rinehart and Winston, Inc. New York (1966), 434 pp. Zbl. 196.456.

[20] R.S.Varga, Matrix iterative analysis, Prentice-Hall, Englewood Cliffs, N.J. (1962), 322 pp. MR 28 ≠≠ 1725, Zbl. 133.86.

[21] J.R.Westlake, A handbook of numerical matrix inversion and solution of linear equations, John Wiley and Sons, New York (1968), 171 pp. MR 36 ≠≠ 4794, Zbl. 155.199.

[22] J.H.Wilkinson, The algebraic eigenvalue problem, Clarendon Press, Oxford (1965), 662 pp. MR 32 ≠≠ 1894, Zbl. 202.154.

[23] D.M.Young, Iterative solution of large linear systems, Academic Press, New York (1971), 570 pp.

Inertia Theory:

[24] B.D.O.Anderson, Application of the second method of Lyapunov to the proof of the Markov stability criterion, Internat. J. Control 5(1967), 473-482. Zbl. 189.460.

[25] S.Avraham and R.Loewy, On the inertia of the Lyapunov transform $AH+HA^*$ for $H > 0$, Linear and Multilinear Algebra 6(1978), 1-21. Zbl. 388.15009.

[26] C.A.Bahl and B.E.Cain, The inertia of diagonal multiples of 3x3 real matrices, Linear Algebra and Appl. 18(1977), 267-280. Zbl. 379.15004.

[27] S.Barnett, Matrices, polynomials, and linear time-invariant systems, IEEE Trans. Automatic Control AC-18(1973), 1-10. [Correction to: "Matrices, polynomials, and linear time-invariant systems", IEEE Trans. Automatic Control AC-18 (1973), 687] MR 56 ## 11160a; 11160b, Zbl.268.93016, 271.93012.

[28] S.Barnett, Some topics in algebraic systems theory: a survey, Internat. J. Control 19(1974), 669-688. MR 49 ## 4603, Zbl. 282.93002.

[29] S.Barnett, Comments on "The matrix sign function and computations in systems" by E.D.Denman and A.N.Beavers, Appl. Math. Comput. 4(1978), 277-279. Zbl. 408.65023.

[30] S.Barnett and D.D.Šiljak, Routh's algorithm: a centennial survey, SIAM Rev. 19(1977), 472-489. MR 56 ## 4985, Zbl. 361.93003.

[31] S.Barnett and C.Storey, Analysis and synthesis of stability matrices, J. Differential Equations 3(1967), 414-422. MR 35 ## 1876, Zbl. 156.36.

[32] A.Berman and A.Ben-Israel, Linear equations over cones with interior: a solvability theorem with applications to matrix theory, Linear Algebra and Appl. 7(1973), 139-149. MR 47 ## 3408, Zbl.254.15010.

[33] B.E.Cain, Inertia theory for operators on a Hilbert space, Ph.D. Thesis, The University of Wisconsin (1968). Dis. Abs. 29B (1968), 676-B.

[34] B.E.Cain, An inertia theory for operators on a Hilbert space, J. Math. Anal. Appl. 41 (1973), 97-114. MR 47 ## 5637, Zbl. 251.47005.

[35] B.E.Cain, The inertial aspects of Stein's condition $H-C^*HC >> 0$, Trans. Amer. Math. Soc. 196(1974), 79-91. MR 50 ## 2941, Zbl. 259.47022, 285.47017.

[36] B.E.Cain, Inertia theory, Linear Algebra and Appl. 30(1980), 211-240.

[37] D.Carlson, Rank and inertia theorems for matrices: the semi-definite case, Ph.D. Thesis, The University of Wisconsin (1963). Dis. Abs. 24(1963), 1627.

[38] D.Carlson, Rank and inertia bounds for matrices under $R(AH) \geq 0$, J. Math. Anal. Appl. 10(1965), 100-111. MR 30 ≠≠ 4770, Zbl. 133-261.

[39] D.Carlson, On real eigenvalues of complex matrices, Pacific J. Math. 15(1965), 1119-1129. MR 32 ≠≠ 5667, Zbl. 142.270.

[40] D.Carlson, A new criterion for H-stability of complex matrices, Linear Algebra and Appl. 1(1968), 59-64. MR 37 ≠≠ 1383, Zbl. 155.63.

[41] D.Carlson and B.N.Datta, The Lyapunov matrix equation $SA+A^*S=S^*B^*BS$, Linear Algebra and Appl. 28(1979), 43-52. Zbl. 422.15010.

[42] D.Carlson and B.N.Datta, On the effective computation of the inertia of a non-Hermitian matrix, Numer. Math. 33(1979), 315-322. Zbl. 402.15003, 414.15012.

[43] D.Carlson and R.Hill, Generalized controllability and inertia theory, Linear Algebra and Appl. 15(1976), 177-187. Zbl. 338.15010.

[44] D.Carlson and R.D.Hill, Controllability and inertia theory for functions of a matrix, J. Math. Anal. Appl. 59(1977), 260-266. MR 56 ≠≠ 11138, Zbl. 359.15009.

[45] D.Carlson and T.L.Markham, Schur complements of diagonally dominant matrices, Czechoslovak Math. J. 29(104) (1979), 246-251. Zbl. 423.15008.

[46] D.Carlson and H.Schneider, Inertia theorems for matrices: the semi-definite case, Bull. Amer. Math. Soc. 68(1962), 479-483. MR 26 ≠≠ 6184, Zbl.192.134.

[47] D.Carlson and H.Schneider, Inertia theorems for matrices: the semidefinite case, J. Math. Anal. Appl. 6(1963), 430-446. MR 26 ≠≠ 6185, Zbl. 192.134.

[48] C.T.Chen, A generalization of the inertia theorem, SIAM J. Appl. Math. 25(1973), 158-161. MR 49 ≠≠ 315, Zbl. 273.15009.

[49] C.T.Chen, Inertia theorem for general matrix equations, J. Math. Anal. Appl. 49(1975), 207-210. MR 50 ≠≠ 4610, Zbl. 351.15010.

[50] R.Datko and V.Seshadri, A characterization and a canonical decomposition of Hurwitzian matrices, Amer. Math. Monthly, 77(1970), 732-733. MR 42 ## 4568, Zbl. 255.15012.

[51] B.N.Datta, An inertia theorem for the Schwarz matrix, IEEE Trans. Automatic Control AC-20(1975), 274. MR 55 ## 367, Zbl. 299.93025.

[52] B.N.Datta, On the Routh-Hurwitz-Fujiwara and the Schur-Cohn-Fujiwara theorems for the root-separation problem, Linear Algebra and Appl. 22(1978), 235-245. Zbl. 387.15011.

[53] B.N.Datta, Two inertia theorems for Hessenberg matrices and their applications to stability analysis of linear control systems, Matrix Tensor Quart. 29(1978), 55-59. Zbl. 399.15009.

[54] E.D.Denman and A.N.Beavers Jr., The matrix sign function and computations in systems, Appl. Math. Comput. 2(1976), 63-94. MR 52 ## 13886, Zbl. 398.65023.

[55] R.J.Duffin, Algorithms for classical stability problems, SIAM Rev. 11(1969), 196-213. MR 40 ## 2981, Zbl.175.98.

[56] A.T.Fuller, Conditions for a matrix to have only characteristic roots with negative real parts, J. Math. Anal. Appl. 23(1968), 71-98. MR 37 ## 4097, Zbl. 157.157.

[57] E.V.Haynsworth, Determination of the inertia of a partitioned Hermitian matrix, Linear Algebra and Appl. 1(1968), 73-81. MR 36 ## 6440, Zbl. 155-63.

[58] E.V.Haynsworth and A.M.Ostrowski, On the inertia of some classes of partitioned matrices, Linear Algebra and Appl. 1(1968), 299-316. MR 38 ## 166, Zbl. 186.337.

[59] R.D.Hill, Generalized inertia theory for complex matrices, Ph.D. Thesis, Oregon State University (1968). Dis. Abs. 29B(1968), 1433-B.

[60] R.D.Hill, Inertia theory for simultaneously triangulable complex matrices, Linear Algebra and Appl. 2(1969), 131-142. MR 39 ## 6902, Zbl. 186.339.

[61] R.D.Hill, Eigenvalue location using certain matrix functions and geometric curves, Linear Algebra and Appl. 16(1977), 83-91. MR 57 ## 6058.

[62] A.S.Householder, Bigradients and the problem of Routh and Hurwitz, SIAM Rev. 10(1968), 56-66. [Erratum: SIAM Rev. 10(1968), 438] MR 37 ## 5371, 38 ## 3412, Zbl. 169.90, 182.92.

[63] J.L.Howland, Matrix equations and the separation of matrix eigenvalues, J. Math. Anal. Appl. 33(1971), 683-691. MR 43 ≠≠ 233, Zbl. 187.301, 207.43.

[64] C.R.Johnson, A Lyapunov theorem for angular cones, J. Res. Nat. Bur. Standards, Sect. B 78B(1974), 7-10. MR 48 ≠≠ 11161, Zbl. 283.15013.

[65] C.R.Johnson, A local Lyapunov theorem and the stability of sums, Linear Algebra and Appl. 13(1976), 37-43. MR 55 ≠≠ 7521, Zbl. 333.15004.

[66] C.R.Johnson, The inertia of a product of two Hermitian matrices, J. Math. Anal. Appl. 57(1977), 85-90. MR 55 ≠≠ 5665, Zbl. 356.15011.

[67] G.T.Joyce and S. Barnett, Remarks on the inertia of a matrix, Linear Algebra and Appl. 3(1970), 1-5. MR 41 ≠≠ 6876, Zbl. 188.78.

[68] R.Loewy, On the Lyapunov transformation for stable matrices, Ph.D. Thesis, California Institute of Technology (1972). Dis. Abs. 33B(1972), 2704-B.

[69] A.M.Ostrowski, A quantitative formulation of Sylvester's law of inertia, Proc. Nat. Acad. Sci. U.S.A. 45(1959), 740-744. MR 22 ≠≠ 1588, Zbl. 87.18.

[70] A.M.Ostrowski, A quantitative formulation of Sylvester's law of inertia II, Proc. Nat. Acad. Sci. U.S.A. 46(1960), 859-862. MR 22 ≠≠ 9504, Zbl. 95.12.

[71] A.Ostrowski and H.Schneider, Some theorems on the inertia of general matrices, J. Math. Anal. Appl. 4(1962), 72-84. MR 26 ≠≠ 124, Zbl. 112.14.

[72] K.J.Palmer, A sufficient condition that a matrix have eigenvalues with non-zero real parts, Linear and Multilinear Algebra 7(1979), 43-45. Zbl. 414.15009.

[73] C.Popeea and L.Lupas, Numerical stabilizability tests by a matrix sign function, 22(1977), 654-656. Zbl. 361.93038.

[74] R.Rado, An extension of Sylvester's law of inertia, Linear Algebra and Appl. 1(1968), 29-31. MR 37 ≠≠ 234, Zbl. 164.33.

[75] R.K.S.Rathore and C.S.K.Chetty, Some angularity and inertia theorems related to normal matrices, Linear Algebra and Appl. (~~in press~~) 40 (1981), 69-77.

[76] W.T.Reid, A matrix equation related to a non-oscillation criterion and Liapunov stability, Quart. Appl. Math. 23(1965), 83-87. MR 31 ## 1263, Zbl. 132.7.

[77] H.Schneider, Positive operators and an inertia theorem, Numer. Math. 7(1965), 11-17. MR 30 ## 3888, Zbl. 158.280.

[78] H.Schneider, Topological aspects of Sylvester's theorem on the inertia of Hermitian matrices, Amer. Math. Monthly, 73(1966), 817-821. MR 34 ## 1335, Zbl. 145.252.

[79] H.R.Schwarz, Ein verfahren zur stabilitätsfrage bei matrizen-eigenwertproblemen, Z. Angew. Math. Phys. 7(1956), 473-500. MR 18 ## 676, Zbl. 73.339.

[80] P.Stein, Some general theorems on iterants, J. Res. Nat. Bur. Standards 48 (1952), 82-83. MR 13 ## 813.

[81] O.Taussky, A remark on a theorem of Lyapunov, J. Math. Anal. Appl. 2(1961), 105-107. MR 23 ## A1649, Zbl.158.282.

[82] O.Taussky, A generalization of a theorem of Lyapunov, J. Soc. Indust. Appl. Math. 9(1961), 640-643. MR 24 ## A3170, Zbl. 108.12.

[83] O.Taussky, Matrices C with $C^n \to 0$, J. Algebra, 1(1964), 5-10. MR 28 ## 5069, Zbl. 126.28.

[84] O.Taussky, On stable matrices, Colloques. internat. centre nat. Rech. Sci. in programmation en mathématiques numériques 165(1968), 75-88. MR 37 ## 6298, Zbl. 208.399.

[85] R.C.Thompson, Inertial properties of eigenvalues, J. Math. Anal. Appl. 41(1973), 192-198. MR 53 ## 485, Zbl. 259.15007.

[86] R.C.Thompson, Inertial properties of eigenvalues II, J. Math. Anal. Appl. 58(1977), 572-577. MR 56 ## 3035, Zbl. 362.15010.

[87] H.S.Wall, Polynomials whose zeros have negative real parts, Amer. Math. Monthly 52(1945), 308-322. MR 7 ## 62, Zbl. 60.55.

[88] H.Wielandt, On the eigenvalues of A+B and AB, J. Res. Nat. Bur. Standards Sect. B 77B(1973), 61-63. MR 49 ## 318, Zbl. 271.15010.

[89] H.K.Wimmer, On the Ostrowski-Schneider inertia theorem, J. Math. Anal. Appl. 41(1973), 164-169. MR 47 ## 6729, Zbl. 251.15011.

[90] H.K.Wimmer, Inertia theorems for matrices, controllability, and linear vibrations, Linear Algebra and Appl. 8(1974), 337-343. MR 52 ≠≠ 15190, Zbl. 288.15015.

[91] H.K.Wimmer, An inertia theorem for tridiagonal matrices and a criterion of Wall on continued fractions, Linear Algebra and Appl. 9(1974), 41-44. MR 50 ≠≠ 13079, Zbl. 294.15009.

[92] H.K.Wimmer, Generalizations of theorems of Lyapunov and Stein, Linear Algebra and Appl. 10(1975), 139-146. MR 50 ≠≠ 9919, Zbl. 307.15002.

[93] H.K.Wimmer and A.D.Ziebur, Remarks on inertia theorems for matrices, Czechoslovak Math. J. 25(100) (1975), 556-561. MR 53 ≠≠ 1938, Zbl. 344.15008.

[94] J.S.W.Wong, A remark on a theorem of Lyapunov, Canad. Math. Bull. 13(1970), 141-143. MR 43 ≠≠ 645, Zbl. 195.102.

Linear Transformations with Invariants:

[95] L.B.Beasley, Linear transformations on matrices: the invariance of certain matrix functions, Ph.D. Thesis, The University of British Columbia (1969). Dis. Abs. 31B(1970), 276-B.

[96] L.B.Beasley, Linear transformations on matrices: the invariance of rank k matrices, Linear Algebra and Appl. 3(1970), 407-427. MR 42 ≠≠ 6010, Zbl. 206.39.

[97] L.B.Beasley, Linear transformations on matrices: the invariance of the third elementary symmetric function, Canad. J. Math. 22(1970), 746-752. MR 42 ≠≠ 3100, Zbl. 231.15012.

[98] L.B.Beasley, Linear transformations on matrices: the invariance of commuting pairs of matrices, Linear and Multilinear Algebra 6(1978), 179-183. Zbl. 397.15010.

[99] D.Benson, A characterization of linear transformations which leave the doubly stochastic matrices invariant, Linear and Multilinear Algebra 6(1978), 65-72. Zbl. 379.15011.

[100] P.Botta, Linear transformations that preserve the permanent, Proc. Amer. Math. Soc. 18(1967), 566-569. MR 35 ≠≠ 4240, Zbl. 148.257.

[101] E.P.Botta, Linear transformations on matrices: the invariance of a class of general matrix functions, Canad. J. Math. 19(1967), 281-290. MR 35 ≠≠ 1609, Zbl. 147.278.

[102] P.Botta, Linear transformations on matrices: the invariance of a class of general matrix functions. II, Canad. J. Math. 20(1968), 739-748. MR 37 ≠≠ 2772, Zbl. 258.15006.

[103] P.Botta, Linear maps that preserve singular and nonsingular matrices, Linear Algebra and Appl. 20(1978), 45-49. MR 58 ≠≠ 5724, Zbl. 371.15005.

[104] P.Botta, Linear transformations preserving the unitary group, Linear and Multilinear Algebra 8(1979), 89-96. Zbl. 435.15004.

[105] E.P.Botta and S.Pierce, The preservers of any orthogonal group, Pacific J. Math. 70(1977), 37-49. Zbl. 381.15004.

[106] M.D.Choi, Completely positive linear maps on complex matrices, Linear Algebra and Appl. 10(1975), 285-290. MR 51 ≠≠ 12901, Zbl. 327.15018.

[107] J.Dieudonné, Sur une généralisation du groupe orthogonal à quatre variables, Arch. Math. 1(1949), 282-287. MR 10-586, Zbl. 32.106.

[108] D.Ž.Djoković, Linear transformations of tensor products preserving a fixed rank, Pacific J. Math. 30(1969), 411-414. MR 40 ≠≠ 2704, Zbl. 185.83.

[109] M.L.Eaton, On linear transformations which preserve the determinant, Illinois J. Math. 13(1969), 722-727. MR 40 ≠≠ 4281, Zbl. 186.55.

[110] R.D.Hill, Linear transformations which preserve Hermitian matrices, Linear Algebra and Appl. 6(1973), 257-262. MR 47 ≠≠ 255, Zbl. 252.15012.

[111] R.Howard, Linear maps that preserve matrices annihilated by a polynomial, Linear Algebra and Appl. 30(1980), 167-176. Zbl. 439.15005.

[112] A.Kovacs, Trace preserving linear transformations on matrix algebras, Linear and Multilinear Algebra 4(1977), 243-250. MR 55 ≠≠ 2959, Zbl. 361.15012.

[113] M.H.Lim, Linear transformations on symmetric matrices, Southeast Asian Bull. Math. 2(1978), 58. Zbl. 426.15009.

[114] M.H.Lim and H.Ong, Linear transformations on symmetric matrices that preserve the permanent, Linear Algebra and Appl. 21(1978), 143-151. Zbl. 384.15005.

[115] M.J.S.Lim, Rank preservers of skew-symmetric matrices, Pacific J.Math. 35(1970), 169-174. MR 43 ≠≠ 6229, Zbl. 223.15014.

[116] M.Marcus, All linear operators leaving the unitary group invariant, Duke Math. J. 26(1959), 155-163. MR 21 ≠≠ 54, Zbl.84.17.

[117] M.Marcus, Linear operations on matrices, Amer. Math. Monthly 69(1962), 837-847. MR 26 ≠≠ 5007, Zbl. 108.11.

[118] M.Marcus, Linear transformations on matrices, J. Res. Nat. Bur. Standards Sect. B 75B(1971), 107-113. MR 46 ≠≠ 9056, Zbl. 244.15013.

[119] M.Marcus and I.Filippenko, Linear operators preserving the decomposable numerical range, Linear and Multilinear Algebra 7(1979), 27-36. Zbl. 399.15013.

[120] M.Marcus and J.Holmes, Groups of linear operators defined by group characters, Trans. Amer. Math. Soc. 172(1972), 177-194. MR 46 ≠≠ 9184, Zbl. 263.20023.

[121] M.Marcus and F.May, On a theorem of I.Schur concerning matrix transformations, Arch. Math. 11(1960), 401-404. MR 24 ≠≠ A134, Zbl. 98.14.

[122] M.Marcus and F.C.May, The permanent function, Canad. J. Math. 14(1962), 177-189. MR 25 ≠≠ 1178, Zbl. 106.16.

[123] M.Marcus and H.Minc, The invariance of symmetric functions of singular values, Pacific J. Math. 12(1962), 327-332. MR 26 ≠≠ 5005, Zbl. 105.10.

[124] M.Marcus and B.N.Moyls, Transformations on tensor product spaces, Pacific J. Math. 9(1959), 1215-1221. MR 21 ≠≠ 7219, Zbl. 89.39.

[125] M.Marcus and B.N.Moyls, Linear transformations on algebras of matrices, Canad. J. Math. 11(1959), 61-66. MR 20 ≠≠ 6432, Zbl. 86.17.

[126] M.Marcus and R.Purves, Linear transformations on algebras of matrices: the invariance of the elementary symmetric functions, Canad. J. Math. 11(1959), 383-396. MR 21 ≠≠ 4167, Zbl. 86.17.

[127] M.Marcus and R.Westwick, Linear maps on skew symmetric matrices: the invariance of elementary symmetric functions, Pacific J. Math. 10(1960), 917-924. MR 22 ≠≠ 5641, Zbl. 93.242.

[128] H.Minc, Linear transformations on nonnegative matrices, Linear Algebra and Appl. 9(1974), 149-153. MR 51 ≠≠ 559, Zbl. 293.15019.

[129] H.Minc, Linear transformations on matrices: rank 1 preservers and determinant preservers, Linear and Multilinear Algebra, 4(1977), 265-272. MR 55 ≠≠ 8060, Zbl. 351.15005.

[130] B.N.Moyls, M.Marcus and H.Minc, Permanent preservers on the space of doubly stochastic matrices, Canad. J. Math. 14(1962), 190-194. MR 25 ≠≠ 95, Zbl.103.252.

[131] H.Ong, Linear transformations on matrices: the invariance of generalized permutation matrices. III, Linear Algebra and Appl. 15(1976), 119-151. Zbl. 369.15010.

[132] H.Ong, Linear transformations on matrices: the invariance of a class of general matrix functions, Canad. J. Math. 29(1977), 937-946. MR 56 ≠≠ 5592, Zbl. 344.15005, 353.15008.

[133] H.Ong and E.P.Botta, Linear transformations on matrices: the invariance of generalized permutation matrices. I, Canad. J. Math. 28(1976), 455-472. MR 53 ≠≠ 13284, Zbl. 359.15018.

[134] V.J.Pellegrini, Numerical range preserving operators on a Banach algebra, Studia Math. 54(1975), 143-147. MR 52 ≠≠ 8941, Zbl. 314.47002.

[135] S.Pierce, Linear operators preserving the real symplectic group, Canad. J. Math. 27(1975), 715-724. MR 52 ≠≠ 3192, Zbl. 284.15003, 303.15003.

[136] S.Pierce, Discriminant preserving linear maps, Linear and Multilinear Algebra 8(1979), 101-114. Zbl. 419.15002.

[137] S.Pierce and W.Watkins, Invariants of linear maps on matric algebras, Linear and Multilinear Algebra 6(1978), 185-200. Zbl. 397.15011.

[138] J. de Pillis, Linear transfomations which preserve Hermitian and positive semidefinite operators, Pacific J. Math. 23(1967), 129-137. MR 36 ≠≠ 5153, Zbl. 166.300

[139] J.A.Poluikis and R.D.Hill, Completely positive and Hermitian-preserving linear transformations, Linear Algebra and Appl. (to appear)

[140] J.Rackusin and W.Watkins, Linear maps on matrices: the preservation of symmetric polynomials, Linear Algebra and Appl. 17(1977), 269-276. Zbl. 371.15004.

[141] E.A.Robinson, Multilinear transformations on matrices, Linear Algebra and Appl. 20(1978), 205-218. Zbl. 376.15014.

[142] S.Russo, Trace preserving mappings of matrix algebras, Duke Math. J. 36(1969), 297-300. MR 39 ## 1501, Zbl. 181.407.

[143] R.Sinkhorn, Linear transformations under which the doubly stochastic matrices are invariant, Proc. Amer. Math. Soc. 27(1971), 213-221. MR 42 ## 4573, Zbl. 211.353.

[144] W.Watkins, Linear maps that preserve commuting pairs of matrices, Linear Algebra and Appl. 14(1976), 29-35. Zbl. 329.15005.

[145] A.Wei, Linear transformations preserving the real orthogonal group, Canad. J. Math. 27(1975), 561-572. MR 52 ## 3193, Zbl. 271.20020, 303.20032.

The Matrix Equation AX+XB=C:

[146] B.D.O.Anderson, Solution of quadratic matrix equations, Electron. Lett. 2(1966), 371-372.

[147] C.Apostol, On the operator equation TX-XV=A, Proc. Amer. Math. Soc. 59(1976), 115-118. MR 54 ## 977, Zbl. 347.47010.

[148] J.K.Baksalary and R.Kala, The matrix equation AX-YB=C, Linear Algebra and Appl. 25(1979), 41-43. Zbl.403.15010.

[149] J.K.Baksalary and R.Kala, The matrix equation AXB+CYD=E, Linear Algebra and Appl. 30(1980), 141-147. Zbl. 437.15005.

[150] G.P.Barker, Normal matrices and the Lyapunov equations, SIAM J. Appl. Math. 26(1974), 1-4. MR 48 ## 8531, Zbl. 243.15009, 273.15015.

[151] G.P.Barker, Common solutions to the Lyapunov equations, Linear Algebra and Appl. 16(1977), 233-235. Zbl. 354.15004.

[152] Y.Bar-Ness and G.Langholz, The solution of the matrix equation XC-BX=D as an eigenvalue problem, Internat. J. Systems Sci. 8(1977), 385-392. MR 56 ≠≠ 4947, Zbl. 353.15022.

[153] S.Barnett, Sensitivity of optimal linear systems to small variations in parameters, Internat. J. Control, 4(1966), 41-48. Zbl. 201.477.

[154] S.Barnett, Remarks on solution of AX+XB=C, Electron. Lett. 7(1971), 383. MR 47 ≠≠ 7904.

[155] S.Barnett, Simplification of the Lyapunov matrix equation $A^TPA - P = -Q$, IEEE Trans. Automatic Control AC-19(1974), 446-447.

[156] S.Barnett, Simplification of certain linear matrix e equations, IEEE Trans. Automatic Control AC-21(1976), 115-116. MR 57 ≠≠ 349, Zbl. 325.15010.

[157] S.Barnett and C.Storey, Stability analysis of constant linear systems of Lyapunov's second method, Electron. Lett. 2(1966), 165-166.

[158] S.Barnett and C.Storey, Solution of the Lyapunov matrix equation, Electron. Lett. 2(1966), 466-467.

[159] S.Barnett and C.Storey, Remarks on numerical solution of the Lyapunov matrix equation, Electron. Lett. 3(1967), 417-418.

[160] S.Barnett and C.Storey, Some applications of the Lyapunov matrix equation, J. Inst. Math. Appl. 4(1968), 33-42. Zbl. 155-129.

[161] A.Y.Barraud, A numerical algorithm to solve $A^TXA-X=Q$, IEEE Trans. Automatic Control AC-22(1977), 883-885. MR 56 ≠≠ 17961, Zbl. 361.65022.

[162] A.Y.Barraud, A new numerical solution of $\dot{X} = A_1X + XA_2 + D$, $X(0)=C$, IEEE Trans. Automatic Control AC-22(1977), 976-977. Zbl. 383.65049.

[163] A.Y.Barraud, Commens on "The numerical solution of $A^TQ+QA=-C$", IEEE Trans. Automatic Control AC-24(1979), 671-672. Zbl. 419.65027.

[164] R.H.Bartels and G.W.Stewart, Algorithm 432 Solution of the matrix equation AX+XB=C, Comm. ACM. 15(1972), 820-826.

[165] A.N.Beavers Jr. and E.D.Denman, A new solution method for quadratic matrix equations, Math. Biosci. 20(1974), 135-143. MR 49 ≠≠ 5032, Zbl. 278.65040.

[166] A.N.Beavers Jr. and E.D.Denman, Asymptotic solutions to the matrix Riccati equation, Math. Biosci. 20(1974), 339-344. MR 51 ≠≠ 14598, Zbl. 282.65067.

[167] A.N.Beavers Jr. and E.D.Denman, A new solution method for the Lyapunov matrix equation, SIAM J. Appl. Math. 29(1975), 416-421. MR 52 ≠≠ 3190, Zbl. 317.15011.

[168] P.R.Bélanger and T.P.McGillivray, Computational experience with the solution of the matrix Lyapunov equation, IEEE Trans. Automatic Control AC-21(1976), 799-800. Zbl. 359.65027.

[169] R.Bellman, Notes on matrix theory-X. A problem in control, Quart. Appl. Math. 14(1957), 417-419. MR 18 ≠≠ 576, Zbl. 78.124.

[170] R.Bellman, Kronecker products and the second method of Lyapunov, Math. Nachr. 20(1959), 17-19. MR 22 ≠≠ 48, Zbl. 87-15.

[171] C.S.Berger, A numerical solution of the matrix equation $P = \phi P \phi^{t} + S$, IEEE Trans. Automatic Control, AC-16(1971), 381-382.

[172] T.A.Bickart, Direct solution method for $A_1E+EA_2=-D$, IEEE Trans. Automatic Control AC-22(1977), 467-468. Zbl. 363.65024.

[173] W.G.Bickley and J.McNamee, Matrix and other difect methods for the solution of systems of linear difference equations, Philos. Trans. Roy. Soc. London Ser. A. 252(1960), 69-131. MR 22 ≠≠ 4897, Zbl. 92.130.

[174] S.P.Bingulac, An alternate approach to expanding $PA+A^{t}P=-Q$, IEEE Trans. Automatic Control AC-15(1970), 135-137.

[175] T.L.Boullion and G.D.Poole, A characterization of the general solution of the matrix equation AX+XB=C, Indust. Math. 20(1970), 91-95. MR 45 ≠≠ 6839.

[176] S.Campbell and J.Daughtry, The stable solutions of quadratic matrix equations, Proc. Amer. Math. Soc. 74(1979), 19-23. Zbl. 403.15012.

[177] C.F.Chen and L.S.Shieh, A note on expanding $PA+A^{T}P=-Q$, IEEE Trans. Automatic Control AC-13(1968), 122-123.

[178] M.R.Chidambara and N.Viswanatham, Some new applications of the solution of the equation $AX+XB=-Q$, J. Indian Inst. Sci. 56(1974), 175-184. MR 51 ## 7664, Zbl. 333.93031.

[179] W.A.Coppel, Matrix quadratic equations, Bull. Austral. Math. Soc. 10(1974), 377-401. MR 51 ## 3623, Zbl. 276.15019.

[180] J.Daughtry, Isolated solutions of quadratic matrix equations, Linear Algebra and Appl. 21(1978), 89-94. MR 58 ## 5720, Zbl. 363.47007, 375.47008.

[181] C.W.Davis Jr., The Lyapunov and Stein transformations and related results, Ph.D. Thesis, Auburn University.(1971). Dis. Abs. 32B(1971), 3490-B.

[182] E.J.Davison, The numerical solution of $\dot{X} = A_1X+XA_2+D$, $X(0)=C$, IEEE Trans. Automatic Control AC-20(1975), 566-567. MR 52 ## 16029, Zbl. 308.65044.

[183] E.J.Davison and F.T.Man, The numerical solution of $A'Q+QA=-C$, IEEE Trans. Automatic Control AC-13(1968), 448-449. MR 38 ## 4010.

[184] A.Dou, Method of undetermined coefficients in linear differential systems and the matrix equation $YB-AY=F$, SIAM J. Appl. Math. 14(1966), 691-696. MR 34 ## 4588, Zbl. 148.65.

[185] W.H.Enright, Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations, ACM Trans. Math. Software 4(1978), 127-136. Zbl. 382.65029.

[186] M.A.Epton, Methods for the solution of $AXD-BXC=E$ and its application in the numerical solution of implicit ordinary differential equations, Nordisk Tidskr Informationbehandling (BIT), 20(1980), 341-345.

[187] R.B.Feinberg, Similarity of partitioned matrices, J. Res. Nat. Bur. Standards Sect. B 79B(1975), 117-125. MR 55 ## 10481, Zbl. 332.15005.

[188] H.Flanders and H.K.Wimmer, On the matrix equations $AX-XB=C$ and $AX-YB=C$, SIAM J. Appl. Math. 32(1977), 707-710. MR 56 ## 5599, Zbl. 385.15008.

[189] H.O.Foulkes, Rational solutions of the matrix equation $XA=BX$, Proc. London Math. Soc.(2)50(1948), 196-209. MR 10 ## 95, Zbl. 30.339.

[190] J.M.Freeman, The tensor product of semigroups and the operator equation SX-XT=A, J. Math. Mech. 19(1970), 819-828. MR 43 ## 7961, Zbl. 206.444.

[191] J.A.Goldstein, On the operator equation AX+XB=Q, Proc. Amer. Math. Soc. 70(1978), 31-34. Zbl. 354.47005, 379.47006.

[192] G.H.Golub, S.Nash and C.Van Loan, A Hessenberg-Schur method for the problem AX+XB=C, IEEE Trans. Automatic Control AC-24(1979), 909-913. Zbl. 421.65022.

[193] D.Gottlieb and M.D.Gunzburger, On the matrix equations $AH+HA^*=A^*H+HA=I$, Linear Algebra and Appl. 17(1977), 277-[illegible]. MR 57 ## 9721, Zbl. 358.15010.

[194] R.J.Guidorzi, Transformation approaches in the solution of the matrix equation $A^TX+XB=P$, IEEE Trans. Automatic Control AC-17(1972), 377-379. Zbl. 259.93025.

[195] W.H.Gustafson, Roth's theorems over commutative rings, Linear Algebra and Appl. 23(1979), 245-251. Zbl. 368.15013.

[196] W.H.Gustafson and J.M.Zelmanowitz, On matrix equivalence and matrix equations, Linear Algebra and Appl. 27(1979), 213-224. Zbl. 419.15009.

[197] I.Hartander, Numerical solution of $A^TS+SA+Q=0$, Information Sci. 4(1972), 35-50. MR 47 ## 1259, Zbl. 229.65040.

[198] R.E.Hartwig, The resultant and the matrix equation AX=XB, SIAM J. Appl. Math. 22(1972), 538-544. MR 46 ## 3528, Zbl. 243.15008.

[199] R.E.Hartwig, Resultants and the solution of AX+XB=-C, SIAM J. Appl. Math. 23(1972), 104-117. MR 46 ## 9067, Zbl. 222.15007, 237.15009.

[200] R.E.Hartwig, AX-XB=C, resultants and generalized inverses, SIAM J. Appl. Math. 28(1975), 154-183. MR 51 ## 8143, Zbl. 265.15004, 294.15006.

[201] R.E.Hartwig, Roth's equivalence problem in unit regular rings, Proc. Amer. Math. Soc. 59(1976), 39-44, MR 53 ## 13297, Zbl. 347.15005.

[202] J.Z.Hearon, Nonsingular solutions of TA-BT=C, Linear Algebra and Appl. 16(1977), 57-63. MT 56 ## 15676, Zbl. 368.15007.

[203] J.A.Heinen, A technique for solving the extended Liapunov matrix equation, Proc. IEEE. 59(1971), 295-296.

[204] J.A.Heinen, A technique for solving the extended discrete Lyapunov matrix equation, IEEE Trans. Automatic Control AC-17(1972), 156-157. Zbl. 262.93028.

[205] W.D.Hoskins, D.S.Meek and D.J.Walton, The numerical solution of the matrix equation XA+AY=F, Nordisk Tidskr. Informationbehandling (BIT), 17(1977), 184-190. Zbl. 358.65025.

[206] W.D.Hoskins, D.S.Meek and D.J.Walton, A rapidly convergent method for the solution of the matrix equation XA+AY=F, J. Comput. Appl. Math. 3(1977), 211-215. MR 57 #f 1846, Zbl. 361. 65025.

[207] W.D.Hoskins, D.S.Meek and D.J.Walton, The numerical solution of $\dot{X} = A_1X+XA_2+D$, $X(0)=C$, IEEE Trans. Automatic Control AC-22(1977), 881-882. MR 56 #f 14900, Zbl. 361.65070.

[208] W.D.Hoskins, D.S.Meek and D.J.Walton, The numerical solution of A'Q+QA=-C, IEEE Trans. Automatic Control AC-22 (1977), 882-883. MR 57 #f 7969. Zbl. 361.65026.

[209] W.D.Hoskins, D.S.Meek and D.J.Walton, High-order iterative methods for the solution of the matrix equation XA+AY=F, Linear Algebra and Appl. 23(1979), 121-139. Zbl. 397.65024.

[210] W.D.Hoskins, G.M.Pathan and D.J.Walton, Solution of bilinear systems arising from high order discretization of Poisson-type equations, Nordisk Tidskr. Informationbehandling (BIT), 20(1980), 212-214. Zbl. 425.65021.

[211] W.D.Hoskins and D.J.Walton, The numerical solution of the time-dependent matrix equation A(t)V(t)+WA(t)=G(t), Internat. J. Systems Sci. 9(1978), 1019-1028. Zbl. 394.65031.

[212] W.D.Hoskins and D.J.Walton, Methods for solving the matrix equation TB+CT=-A and its generalizations, Linear Algebra and Appl. 23(1979), 217-225. Zbl. 401.15012.

[213] J.L.Howland and J.A.Senez, A constructive method for the solution of the stability problem, Numer. Math. 16(1970), 1-7. MR 42 #f 8673. Zbl. 187.38, 197.430.

[214] M.H.Ingraham and H.C.Trimble, On the matric equation TA=BT+C, Amer. J. Math. 63(1941), 9-28. MR 2 ## 243, Zbl. 24.246.

[215] A.Jameson, Solution of the equation AX+XB=C by inversion of an MxM or NxN matrix, SIAM J. Appl. Math. 16(1968), 1020-1023. MR 38 ## 3286, Zbl. 169.352.

[216] C.R.Johnson and M.Newman, A condition for the diagonalizability of a partitioned matrix, J. Res. Nat. Bur. Standards Sect. B. 79B(1975), 45-48. MR 52 ## 8156, Zbl. 334.15007.

[217] E.L.Jones, A reformulation of the algebraic Riccati equation problem, IEEE Trans. Automatic Control AC-21 (1976), 113-114. Zbl. 317.49050.

[218] J.Jones Jr., A diophantine matrix equation, Amer. Math. Monthly, 62(1955), 244-247. MR 16-894. Zbl. 65-248.

[219] J.Jones Jr., On the Lyapunov stability criteria, J. Soc. Indust. Appl. Math. 13(1965), 941-945. MR 32 ## 4136. Zbl. 136.4.

[220] J.Jones Jr., Solution of certain matrix equations, Proc. Amer. Math. Soc. 31(1972), 333-339. MR 45 ## 1945, Zbl. 242.15007.

[221] J.Jones Jr., Explicit solutions of the matrix equation AX-XB=C, Rend. Circ. Mat.Palermo. II Ser. 23(1974), 245-257. MR 53 ## 2983. Zbl. 328.15005.

[222] D.G.Kabe, Solutions of systems of linear symmetric matrix equations, Indust. Math. 29(1979), 9-16. Zbl. 416.15008.

[223] R.E.Kalman and J.E.Bertram, Control system analysis and design via the "second method" of Lyapunov. I. Continuous-time systems, Trans. ASME. Ser. D.J. Basic Engrg. 82(1960), 371-393. MR 28 ## 1039.

[224] R.E.Kalman and R.S.Bucy, New results in linear filtering and prediction theory, Trans. ASME. Ser. D. J.Basic Engrg. 83(1961), 95-108. MR 38 ## 3076.

[225] G.Kitagawa, An algorithm for solving the matrix equation $X=FXF^T+S$, Internat. J. Control. 25(1977), 745-753. Zbl. 364.65025.

[226] D.L.Kleinman, On an iterative technique for Riccati equation computations, IEEE Trans. Automatic Control AC-13(1968), 114-115.

[227] G.Kreisselmeier, A solution of the bilinear matrix equation AY+YB=-Q, SIAM J. Appl. Math. 23(1972), 334-338. MR 47 ## 5025, Zbl. 256.15009.

[228] V.Kučera, The matrix equation AX+XB=C, SIAM J. Appl. Math. 26(1974), 15-25. MR 49 ## 5035. Zbl. 245.15004, 273.15012.

[229] W.H.Kwon and A.E.Pearson, A note on the algebraic matrix Riccati eugation, IEEE Trans. Automatic Control AC-22(1977), 143-144. MR 56 ## 2563, Zbl. 346.93029.

[230] R.T.Lacoss and A.F.Shakal, More $A_1E+EA_2=-D$ and $\dot{X}=A_1X+XA_2+D$, $X(0)=C$, IEEE Trans. Automatic Control AC-21(1976), 405-406. MR 54 ## 11788, Zbl. 359.65069.

[231] P.Lancaster, Explicit solutions of linear matrix equations, SIAM Rev. 12(1970), 544-566. MR 43 ## 4841, Zbl.209.65.

[232] A.J.Laub, A Schur method for solving algebraic Riccati equations, IEEE Trans. Automatic Control AC-24(1979), 913-921. Zbl. 424.65013.

[233] V.Lovass-Nagy and D.L.Powers, On least squares solutions of an inconsistent singular equation AX+XB=C, SIAM J. Appl. Math. 31(1976), 84-88. MR 53 ## 14890, Zbl. 336.15009.

[234] C.S.Lu, Solution of the matrix equation AX+XB=C, Electron. Lett. 7(1971), 185-186. MR 47 ## 7903.

[235] D.G.Luenberger, Invertible solutions to be operator equation TA-BT=C, Proc. Amer, Math. Soc. 16(1965), 1226-1229. MR 32 ## 1562, Zbl. 138.78.

[236] D.G.Luenberger, Observers for multivariable systems, IEEE Trans. Automatic Control AC-11(1966), 190-197.

[237] G.Lumer and M.Rosenblum, Linear operator equations, Proc. Amer. Math. Soc. 10(1959), 32-41. MR 21 ## 2927, Zbl. 133.79.

[238] E.C.Ma, A matrix analysis of beam gridworks, Ph.D. Dissertation, Kansas State University (1962). Dis. Abs. 23(1962), 983.

[239] E.C.Ma, A finite series solution of the matrix equation AX-XB=C, SIAM J. Appl. Math. 14(1966), 490-495. MR 34 ## 1340, Zbl.144.270.

[240] A.G.J.MacFarlane, The calculation of functionals of the time and frequency response of a linear constant coefficient dynamical system, Quart. J. Mech. Appl. Math. 16(1963), 259-271. MR 27 ## 5993, Zbl. 135.130.

[241] F.T.Man, A high-order method of solution for the Lyapunov matrix equation, Comput. J. 14(1971), 291-292. MR 46 ≠≠ 1065, Zbl. 224.65009.

[242] K.Martensson, On the matrix Riccati equation, Information Sci. 3(1971), 17-49. Zbl.206.456.

[243] H.B.Meyer, The matrix equation AZ+B-ZCZ-ZD=0, SIAM J. Appl. Math. 30(1976), 136-142. MR 52 ≠≠ 10768, Zbl. 353.15023.

[244] H.B.Meyer, Matrix Riccati solutions, Linear Algebra and Appl. 20(1978), 131-146. Zbl.383.15012.

[245] R.Meyer-Spasche, A constructive method of solving Liapunov equation for complex matrices, Numer. Math. 19(1972), 433-438, MR 47 ≠≠ 7905, Zbl. 229.65039, 237.65030.

[246] R.Meyer-Spasche, A method of solving the stability problem for complex matrices, Numer. Math. 20(1973) 364-371, MR 49 ≠≠ 6586. Zbl. 241.65036, 249.65023.

[247] B.E.A. Milani, Decomposition of large-scale matrix equations XA+BX=C, Comptes Reundus Proceedings, Seventh Canadian Congress of Applied Mechanics, Universite de Sherbrooke, Canada, pp.911-912.

[248] S.K.Mitra, The matrix equation AXB+CXD=E, SIAM J. Appl. Math. 32 (1977), 823-825. MR 55 ≠≠ 5662, Zbl. 392.15005.

[249] B.P.Molinari, Algebraic solution of matrix linear equations in control theory, Proc. Inst. Elec. Engrs. 116(1969), 1748-1754. MR 46 ≠≠ 4695.

[250] P.C.Müller, Solution of the matrix equations AX+XB=-Q and $S^TX+XS=-Q$, SIAM J. Appl. Math. 18(1970), 682-687. MR 41 ≠≠ 4786, Zbl. 205.51.

[251] H.Neudecker, A note on Kronecker matrix products and matrix equation systems, SIAM J. Appl. Math. 17(1969), 603-606, MR 40 ≠≠ 1414, Zbl. 185.82.

[252] I.S.Pace and S.Barnett, Comparison of numerical methods for solving Liapunov matrix equations, Internat. J. Control 15(1972), 907-915. MR 47 ≠≠ 1261. Zbl. 235 235.65024.

[253] W.V.Parker, The matrix equation AX=XB, Duke Math. J. 17(1950), 43-51, MR 11 ≠≠ 412, Zbl. 41.154.

[254] P.C.Parks, Stability analysis for linear and non-linear systems using Liapunov's second method, Progress in Control Engg. 2 (1964), 31-64. Zbl. 132.326.

[255] J.E.Potter, Matrix quadratic solutions, SIAM J. Appl. Math. 14(1966), 496-501, MR 34 #≠ 1341. Zbl. 144.20.

[256] H.M.Power, Solution of Lyapunov matrix equation for continuous systems via Schwarz and Routh canonical forms, Electron. Lett. 3(1967), 81-82.

[257] H.M.Power, Equivalence of Lyapunov matrix equations for continuous and discrete systems, Electron. Lett. 3(1967), 83.

[258] H.M.Power, A note on the matrix equation $A'LA-L = -K$, IEEE Trans. Automatic Control, AC-14(1969), 411-412. MR 41 #≠ 235.

[259] D.L.Powers, Solving AX+XB=C by control and tearing, Internat. J. Control, 23(1976), 421-425. Zbl. 323.15004.

[260] M.Rosenblum, On the operator equation BX-XA=Q, Duke Math. J. 23(1956), 263-269. MR 18-54, Zbl. 73.330.

[261] M.Rosenblum, The operator equation BX-XA=Q with selfadjoint A and B, Proc. Amer. Math. Soc. 20(1969), 115-120. MR 38 #≠ 1537, Zbl. 167.428.

[262] W.E.Roth, On direct product matrices, Bull. Amer. Math. Soc. 40(1934), 461-468, Zbl. 9.290.

[263] W.E.Roth, The equations AX-YB=C and AX-XB=C in matrices, Proc. Amer. Math. Soc. 3(1952), 392-396. MR 13 #≠ 900, Zbl. 47.19.

[264] D.Rothschild and A.Jameson, Comparison of four numerical algorithms for solving the Liapunov matrix equation, Internat. J. Control. 11(1970), 181-198. Zbl.185.401.

[265] P.Scobey and D.G.Kabe, On linear matrix equations, Canad. Math. Bull. 23(1980), 43-49.

[266] V.E.Šestopal, Solution of the matrix equation AX-XB=C, Math. Notes, 19(1976), 275-276. MR 53 #≠ 10832. Zbl. 334.15008, 344.15010.

[267] P.G.Smith, Numerical solution of the matrix equation $AX+XA^T+B=0$, IEEE Trans. Automatic. Control, AC-16(1971) 278-279.

[268] R.A.Smith, Bounds for quadratic Lyapunov functions, J. Math. Anal. Appl. 12(1965), 425-435. MR 32 #≠ 7887, Zbl. 135.298.

[269] R.A.Smith, Matrix calculations for Liapunov quadratic forms, J. Differential Equations 2(1966), 208-217. MR 32 ≠≠ 5995, Zbl. 151.22.

[270] R.A.Smith, Matrix equation XA+BX=C, SIAM J. Appl. Math. 16(1968), 198-201. MR 36 ≠≠ 7312, Zbl. 157-226.

[271] J.Snyders and M.Zakai, On nonnegative solutions of the equation AD+DA′=-C, SIAM J. Appl. Math. 18(1970), 704-714. MR 41 ≠≠ 3501, Zbl. 203.334.

[272] J.M.Varah, On the separation of two matrices, SIAM J. Numer. Anal. 16(1979), 216-222. Zbl. 435.65034.

[273] W.J.Vetter, Vector structures and solutions of linear matrix equations, Linear Algebra and Appl. 10(1975), 181-188. MR 51 ≠≠ 553, Zbl. 307.15003.

[274] D.J.Walton, Some new methods for the solution of matrix equations arising from discretized partial differential equations, Ph.D, Dissertation, The University of Manitoba (1978). Dis. Abs. 38B (1978), 5480-B.

[275] H.K.Wimmer, On the algebraic Riccati equation Bull. Austral. Math. Soc. 14(1976), 457-461. MR 54 ≠≠ 10298, Zbl. 344.15007.

[276] H.K.Wimmer and A.D.Ziebur, Solving the matrix equation $\Sigma_{\rho=1}^{r} f_\rho(A)Xg_\rho(B)=C$, SIAM Rev. 14(1972), 318-323. MR 46 ≠≠ 7265, Zbl. 244.15006.

[277] N.J.Young, Formulae for the solution of Lyapunov matrix equations, Internat. J. Control. 31(1980), 159-179. Zbl. 432.93050.

[278] I.E.Ziedan, Explicit solution of the Lyapunov-matrix equation, MR 55 ≠≠ 12747. Zbl. 259.93028. IEEE. Trans. Automatic cont AC-17 (1972), 379

Miscellaneous:

[279] C.M.Ablow and J.L.Brenner, Roots and canonical forms for circulant matrices, Trans. Amer. Math. Soc. 107(1963), 360-376. MR 27 ≠≠ 5775, Zbl. 112.250.

[280] C.S.Ballantine, Products of positive definite matrices.IV, Linear Algebra and Appl. 3(1970), 79-114. MR 41 ≠≠ 1766.

[281] C.S.Ballantine, Numerical range of a matrix: some effective criteria, Linear Algebra and Appl. 19(1978), 117-188. MR 58 ≠≠ 8202, Zbl. 381.15010.

[282] A.Buckley, On the solution of certain skew symmetric linear systems, SIAM J. Numer. Anal. 14(1977), 566-570. MR 55 ≠≠ 6803, Zbl. 362.65023.

[283] J.L.Brenner, g-circulant matrices over a field of prime characteristic, Illinois J. Math. 7(1963), 174-179. MR 26 ≠≠ 2453, Zbl. 109.246.

[284] R.Chalkley, Circulant matrices and algebraic equations, Math. Mag. 48(1975), 73-80, MR 50 ≠≠ 13076.

[285] C.Y.Chao, A remark on cyclic matrices, Linear Algebra and Appl. 3(1970), 165-172. MR 42 ≠≠ 3098, Zbl.205.47.

[286] C.G.Cullen, A note on normal matrices, Amer. Math. Monthly, 72(1965), 643-644. MR 31 ≠≠ 1262.

[287] P.J. Davis, Cyclic transformations of polygons and the generalized inverse, Canad. J. Math. 29(1977), 756-770. Zbl. 368.15013.

[288] P.J.Davis, Cyclic transformations of n-gons and related quadratic forms, Linear Algebra and Appl. 25(1979), 57-75. Zbl. 408.15017.

[289] D.J.Evans, On the solution of certain Toeplitz tridiagonal linear systems, SIAM J. Numer. Anal. 17(1980), 675-680.

[290] G.E.Forsythe, Solving linear algebraic equations can be interesting, Bull. Amer. Math. Soc. 59(1953), 299-329. MR 15 ≠≠65 .

[291] I.J.Good, On the inversion of circulant matrices, Biometrika, 37(1950), 185-186. MR 12 ≠≠ 538, Zbl. 37.145.

[292] W.B.Gragg, On Hadamard's theory of polar singularities in Pade Approximants and their applications (Ed. P.R. Graves Morris), Acadamic Press, London (1973).pp.117-123.

[293] R.M.Gray, On the asymptotic eigenvalue distribution of Toeplitz matrices, IEEE Trans. Information Theorey IT-18(1972), 725-730. Zbl. 246.94002.

[294] M.R.Hestenes and M.L.Stein, The solution of linear equations by minimization, NAML Report 52-45, National Bureau of Standards, Los Angeles, 1951.

[295] A.K.Jain, Fast inversion of banded Toeplitz matrices by circular decompositions, IEEE Trans. Acoust. Speech Signal process ASSP-26(1978), 121-126. Zbl. 429.65026.

[296] S.Kaczmarz, Angenäherte Auflösung von systemen linearer Gleichungen, Bull. Acad. Polon. Sciences et Lettres A(1937), 355-357.

[297] P.Laasonen, On the iterative solution of the matrix equation $AX^2-I=0$, MTAC, 12(1958), 109-116, MR.20 ## 5551, Zbl. 83.117.

[298] H.W.Milnes and R.B.Potta, Boundary contraction solution of Laplaces differential equation, J. Assoc. Comput. Mach. 6(1959), 226-235. MR 21 ## 4554, Zbl.88.340.

[299] P.A.Nekrasov, Determination of the unknowns by the method of least squares for very many unknowns (in Russian) Matematiceskii Sbornik, 12(1884), 189-204.

[300] R.k.S.Rathore, The structure of variation preserving operators, Journal of the Indian Math. Soc. 43(1979), 175-186.

[301] R.K.S.Rathore, A study of the method of residual projections with an application to general stationary least squares schemes (communicated to Numer. Math.)

[302] P.A.Roebuck,and S.Barnett, A survey of Toeplitz and and related matrices, Internat. J. systems Sci. 9(1978), 921-934. Zbl. 385.15010.

[303] P.Schuster, K.Sigmund and R.Wolff, Dynamical systems under constant organization I. Topological analysis of a family of non-linear differential equations - a model for catalytic hypercycles, Bull. Math. Biology, 40(1978), 743-769. Zbl. 384.34028.

[304] S.R.Searle, On inverting circulant matrices, Linear Algebra and Appl. 25(1979), 77-89.

[305] K.Tanabe, Projection method for solving a singular system of linear equations and its applications, Numer. Math. 17(1971), 203-214. MR 45 ## 2-00

[306] R.S.Varga, Eigenvalues of circulant matrices,,Pacific J. Math. 4(1954), 151-160. MR 15 ## 745. Zbl.55.10.

[307] N.A.Wiegmann, Normal products of matrices, Duke Math. J. 15(1948), 633-638. MR 10## 230, Zbl. 31.243.

[308] R.Yarlagadda and B.N.Suresh Babu, A note on the application of FFT to the solution of a system of Toeplitz normal equations, IEEE Trans. Circuits and Systems CAS-27(1980), 151-154. Zbl.425.65019.